
> IS 2024
Informacijska družba
Zbornik 27. mednarodne multikonference
Zvezek A
Information Society
Proceedings of the 27th International Multiconference
Volume A
Slovenska konferenca o umetni inteligenci
Slovenian Conference on Artificial Intelligence
Uredniki > Editors:
Mitja Luštrek, Matjaž Gams, Rok Piltaver
10.–11. oktober 2024 > Ljubljana, Slovenija / 10–11 October 2024 > Ljubljana, Slovenia

Zbornik 27. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2024

Zvezek A

Proceedings of the 27th International Multiconference

INFORMATION SOCIETY – IS 2024

Volume A

Slovenska konferenca o umetni inteligenci
Slovenian Conference on Artificial Intelligence

Uredniki / Editors

Mitja Luštrek, Matjaž Gams, Rok Piltaver

http://is.ijs.si

10.–11. oktober 2024 / 10–11 October 2024
Ljubljana, Slovenia

Uredniki:

Mitja Luštrek
Odsek za inteligentne sisteme, Institut »Jožef Stefan«, Ljubljana

Matjaž Gams
Odsek za inteligentne sisteme, Institut »Jožef Stefan«, Ljubljana

Rok Piltaver
Outfit7, Ljubljana

Založnik: Institut »Jožef Stefan«, Ljubljana
Priprava zbornika: Mitja Lasič, Vesna Lasič, Lana Zemljak
Oblikovanje naslovnice: Vesna Lasič

Dostop do e-publikacije:
http://library.ijs.si/Stacks/Proceedings/InformationSociety

Ljubljana, oktober 2024

Informacijska družba
ISSN 2630-371X

Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani
COBISS.SI-ID 214409987
ISBN 978-961-264-299-0 (PDF)

PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2024

Leto 2024 je hkrati udarno in tradicionalno. Že sedaj, še bolj pa v prihodnosti bosta računalništvo, informatika (RI) in umetna inteligenca (UI) igrali ključno vlogo pri oblikovanju napredne in trajnostne družbe. Smo na pragu nove dobe, v kateri generativna umetna inteligenca, kot je ChatGPT, in drugi inovativni pristopi utirajo pot k superinteligenci in singularnosti, ključnim elementom, ki bodo definirali razcvet človeške civilizacije. Naša konferenca je zato hkrati tradicionalna znanstvena, pa tudi povsem akademsko odprta za nove pogumne ideje, inkubator novih pogledov in idej.

Letošnja konferenca ne le da analizira področja RI, temveč prinaša tudi osrednje razprave o perečih temah današnjega časa – ohranjanje okolja, demografski izzivi, zdravstvo in preobrazba družbenih struktur. Razvoj UI ponuja rešitve za skoraj vse izzive, s katerimi se soočamo, kar poudarja pomen sodelovanja med strokovnjaki, raziskovalci in odločevalci, da bi skupaj oblikovali strategije za prihodnost. Zavedamo se, da živimo v času velikih sprememb, kjer je ključno, da s poglobljenim znanjem in inovativnimi pristopi oblikujemo informacijsko družbo, ki bo varna, vključujoča in trajnostna.

Letos smo ponosni, da smo v okviru multikonference združili dvanajst izjemnih konferenc, ki odražajo širino in globino informacijskih ved: CHATMED v zdravstvu, Demografske in družinske analize, Digitalna preobrazba zdravstvene nege, Digitalna vključenost v informacijski družbi – DIGIN 2024, Kognitivna znanost, Konferenca o zdravi dolgoživosti, Legende računalništva in informatike, Mednarodna konferenca o prenosu tehnologij, Miti in resnice o varovanju okolja, Odkrivanje znanja in podatkovna skladišča – SIKDD 2024, Slovenska konferenca o umetni inteligenci, Vzgoja in izobraževanje v RI.

Poleg referatov bodo razprave na okroglih mizah in delavnicah omogočile poglobljeno izmenjavo mnenj, ki bo oblikovala prihodnjo informacijsko družbo. “Legende računalništva in informatike” predstavljajo slovenski “Hall of Fame” za odlične posameznike s tega področja, razširjeni referati, objavljeni v reviji *Informatica* z 48-letno tradicijo odličnosti, in sodelovanje s številnimi akademskimi institucijami in združenji, kot so ACM Slovenija, SLAIS in Inženirska akademija Slovenije, bodo še naprej spodbujali razvoj informacijske družbe. Skupaj bomo gradili temelje za prihodnost, ki bo oblikovana s tehnologijami, osredotočena na človeka in njegove potrebe.

S podelitvijo nagrad, še posebej z nagrado Michie-Turing, se avtonomna RI stroka vsakoletno opredeli do najbolj izstopajočih dosežkov. Nagrado Michie-Turing za izjemen življenjski prispevek k razvoju in promociji informacijske družbe je prejel prof. dr. Borut Žalik. Priznanje za dosežek leta pripada prof. dr. Sašu Džeroskemu za izjemne raziskovalne dosežke. »Informacijsko limono« za najmanj primerno informacijsko tematiko je prejela nabava in razdeljevanjem osebnih računalnikov ministrstva, »informacijsko jagodo« kot najboljšo potezo pa so sprejeli organizatorji tekmovanja ACM Slovenija. Čestitke nagrajencem!

Naša vizija je jasna: prepoznati, izkoristiti in oblikovati priložnosti, ki jih prinaša digitalna preobrazba, ter ustvariti informacijsko družbo, ki bo koristila vsem njenim članom. Vsem sodelujočim se zahvaljujemo za njihov prispevek k tej viziji in se veselimo prihodnjih dosežkov, ki jih bo oblikovala ta konferenca.

Mojca Ciglarič, predsednica programskega odbora

Matjaž Gams, predsednik organizacijskega odbora

PREFACE TO THE MULTICONFERENCE INFORMATION SOCIETY 2024

The year 2024 is both ground-breaking and traditional. Now, and even more so in the future, computer science, informatics (CS/I), and artificial intelligence (AI) will play a crucial role in shaping an advanced and sustainable society. We are on the brink of a new era where generative artificial intelligence, such as ChatGPT, and other innovative approaches are paving the way for superintelligence and singularity—key elements that will define the flourishing of human civilization. Our conference is therefore both a traditional scientific gathering and an academically open incubator for bold new ideas and perspectives.

This year's conference analyzes key CS/I areas and brings forward central discussions on pressing contemporary issues—environmental preservation, demographic challenges, healthcare, and the transformation of social structures. AI development offers solutions to nearly all challenges we face, emphasizing the importance of collaboration between experts, researchers, and policymakers to shape future strategies collectively. We recognize that we live in times of significant change, where it is crucial to build an information society that is safe, inclusive, and sustainable, through deep knowledge and innovative approaches.

This year, we are proud to have brought together twelve exceptional conferences within the multiconference framework, reflecting the breadth and depth of information sciences:

- CHATMED in Healthcare
- Demographic and Family Analyses
- Digital Transformation of Healthcare Nursing
- Digital Inclusion in the Information Society – DIGIN 2024
- Cognitive Science
- Conference on Healthy Longevity
- Legends of Computer Science and Informatics
- International Conference on Technology Transfer
- Myths and Facts on Environmental Protection
- Data Mining and Data Warehouses – SIKDD 2024
- Slovenian Conference on Artificial Intelligence
- Education and Training in CS/IS.

In addition to papers, roundtable discussions and workshops will facilitate in-depth exchanges that will help shape the future information society. The “Legends of Computer Science and Informatics” represents Slovenia’s “Hall of Fame” for outstanding individuals in this field. At the same time, extended papers published in the Informatica journal, with over 48 years of excellence, and collaboration with numerous academic institutions and associations, such as ACM Slovenia, SLAIS, and the Slovenian Academy of Engineering, will continue to foster the development of the information society. Together, we will build the foundation for a future shaped by technology, yet focused on human needs.

The autonomous CS/IS community annually recognizes the most outstanding achievements through the awards ceremony. The Michie-Turing Award for an exceptional lifetime contribution to the development and promotion of the information society was awarded to Prof. Dr. Borut Žalik. The Achievement of the Year Award goes to Prof. Dr. Sašo Džeroski. The "Information Lemon" for the least appropriate information topic was given to the ministry's procurement and distribution of personal computers. At the same time, the "Information Strawberry" for the best initiative was awarded to the organizers of the ACM Slovenia competition. Congratulations to all the award winners!

Our vision is clear: to recognize, seize, and shape the opportunities brought by digital transformation and create an information society that benefits all its members. We thank all participants for their contributions and look forward to this conference's future achievements.

Mojca Ciglarič, Chair of the Program Committee

Matjaž Gams, Chair of the Organizing Committee

KONFERENČNI ODBORI
CONFERENCE COMMITTEES

International Programme Committee

Vladimir Bajic, South Africa
Heiner Benking, Germany
Se Woo Cheon, South Korea
Howie Firth, UK
Olga Fomichova, Russia
Vladimir Fomichov, Russia
Vesna Hljuz Dobric, Croatia
Alfred Inselberg, Israel
Jay Liebowitz, USA
Huan Liu, Singapore
Henz Martin, Germany
Marcin Paprzycki, USA
Claude Sammut, Australia
Jiri Wiedermann, Czech Republic
Xindong Wu, USA
Yiming Ye, USA
Ning Zhong, USA
Wray Buntine, Australia
Bezalel Gavish, USA
Gal A. Kaminka, Israel
Mike Bain, Australia
Michela Milano, Italy
Derong Liu, Chicago, USA
Toby Walsh, Australia
Sergio Campos-Cordobes, Spain
Shabnam Farahmand, Finland
Sergio Crovella, Italy

Organizing Committee

Matjaž Gams, chair
Mitja Luštrek
Lana Zemljak
Vesna Koricki
Mitja Lasič
Blaž Mahnič

Programme Committee

Mojca Ciglarič, chair
Bojan Orel
Franc Solina
Viljan Mahnič
Cene Bavec
Tomaž Kalin
Jozsef Györkös
Tadej Bajd
Jaroslav Berce
Mojca Bernik
Marko Bohanec
Ivan Bratko
Andrej Brodnik
Dušan Caf
Saša Divjak
Tomaž Erjavec
Bogdan Filipič
Andrej Gams
Matjaž Gams
Mitja Luštrek
Marko Grobelnik
Nikola Guid
Marjan Heričko
Borka Jerman Blažič Džonova
Gorazd Kandus
Urban Kordeš
Marjan Krisper
Andrej Kuščer
Jadran Lenarčič
Borut Likar
Janez Malačič
Olga Markič
Dunja Mladenič
Franc Novak
Vladislav Rajkovič
Grega Repovš
Ivan Rozman
Niko Schlamberger
Stanko Strmčnik
Jurij Šilc
Jurij Tasič
Denis Trček
Andrej Ule
Boštjan Vilfan
Baldomir Zajc
Blaž Zupan
Boris Žemva
Leon Žlajpah
Niko Zimic
Rok Piltaver
Toma Strle
Tine Kolenik
Franci Pivec
Uroš Rajkovič
Borut Batagelj
Tomaž Ogrin
Aleš Ude
Bojan Blažica
Matjaž Kljun
Robert Blatnik
Erik Dovgan
Špela Stres
Anton Gradišek

KAZALO / TABLE OF CONTENTS

Slovenska konferenca o umetni inteligenci / Slovenian Conference on Artificial Intelligence ***1***

PREDGOVOR / FOREWORD 3

PROGRAMSKI ODBORI / PROGRAMME COMMITTEES 5

PandaChat-RAG: Towards the Benchmark for Slovenian RAG Applications / Kuzman Taja, Pavleska Tanja, Rupnik Urban, Cigoj Primož 7

Choosing Features for Stress Prediction with Machine Learning / Bengeri Katja, Lukan Junoš, Luštrek Mitja 11

Predictive Modeling of Football Results in the WWIN League of Bosnia and Herzegovina / Vladić Ervin, Mehanović Dželila, Avdić Elma 15

Sarcasm Detection in a Less-Resourced Language / Đoković Lazar, Robnik-Šikonja Marko 19

Speech-to-Service: Using LLMs to Facilitate Recording of Services in Healthcare / Smerkol Maj, Susič Rok, Ratajec Mariša, Halbwachs Helena, Gradišek Anton 23

Performance Comparison of Axle Weight Prediction Algorithms on Time-Series Data / Kolar Žiga, Susič David, Konečnik Martin, Prestor Domen, Pejanovič Nosaka Tomo, Kulauzović Bajko, Kalin Jan, Skobir Matjaž, Gams Matjaž 27

Comparison of Feature- and Embedding-based Approaches for Audio and Visual Emotion Classification / Trojer Sebastijan, Anžur Zoja, Luštrek Mitja, Slapničar Gašper 31

Multi-modal Data Collection and Preliminary Statistical Analysis for Cognitive Load Assessment / Krstevska Ana, Kramar Sebastjan, Gjoreski Hristijan, Gjoreski Martin, Lukan Junoš, Trojer Sebastijan, Luštrek Mitja, Slapničar Gašper 35

Predicting Health-Related Absenteeism with Machine Learning: A Case Study / Piciga Aleksander, Kukar Matjaž 39

Puzzle Generation for Ultimate-Tic-Tac-Toe / Zirkelbach Maj, Sadikov Aleksander 43

Ethical Consideration and Sociological Challenges in the Integration of Artificial Intelligence in Mental Health Services / Poljak Lukek Saša 47

Optimization Problem Inspector: A Tool for Analysis of Industrial Optimization Problems and Their Solutions / Tušar Tea, Cork Jordan, Andova Andrejaana, Filipič Bogdan 51

Multi-Agent System for Autonomous Table Football: A Winning Strategy / Založnik Marcel, Šoln Kristjan 55

Towards a Decision Support System for Project Planning: Multi-Criteria Evaluation of Past Projects Success / Hafner Miha, Bohanec Marko 59

Minimizing Costs and Risks in Demand Response Optimization: Insights from Initial Experiments / Nedić Mila, Tušar Tea 63

Predicting Hydrogen Adsorption Energies on Platinum Nanoparticles and Surfaces With Machine Learning / Gašparič Lea, Kokalj Anton, Džeroski Sašo 67

SmartCHANGE Risk Prediction Tool: Demonstrating Risk Assessment for Children and Youth / Jordan Marko, Reščič Nina, Kramar Sebastjan, Založnik Marcel, Luštrek Mitja 71

Predicting Mental States During VR Sessions Using Sensor Data and Machine Learning / Kizhevska Emilija, Luštrek Mitja 75

Biomarker Prediction in Colorectal Cancer Using Multiple Instance Learning / Shulajkovska Miljana, Jelenc Matej, Jonnagaddala Jitenndra, Gradišek Anton 79

Feature-Based Emotion Classification Using Eye-Tracking Data / Božak Tomi, Luštrek Mitja, Slapničar Gašper 83

Indeks avtorjev / Author index ***87***

Zbornik 27. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2024

Zvezek A

Proceedings of the 27th International Multiconference

INFORMATION SOCIETY – IS 2024

Volume A

Slovenska konferenca o umetni inteligenci
Slovenian Conference on Artificial Intelligence

Uredniki / Editors

Mitja Luštrek, Matjaž Gams, Rok Piltaver

http://is.ijs.si

10.–11. oktober 2024 / 10–11 October 2024
Ljubljana, Slovenia

PREDGOVOR

Umetna inteligenca doživlja neverjeten in pospešen razvoj, ko se po tričetrt stoletja, ko je Alan Mathison Turing postavil temelje računalništva in umetne inteligence, končno približuje ne le človeški inteligenci, temveč tudi drugim ključnim človeškim lastnostim, kot sta ustvarjalnost, čustvena inteligenca in zavest. Na številnih področjih umetna inteligenca že presega zmogljivosti večine ljudi in celo strokovnjakov. Veliki jezikovni modeli dosegajo tovrstne rezultate tudi pri dosti manj strukturiranih problemih, kot je bilo predstavljivo pred nekaj leti, npr. pri strokovnih izpitih ter besedilnih nalogah iz matematike in programiranja.

Generativna umetna inteligenca že zdaj spreminja svet. Postala je nepogrešljivo orodje v poslovnem svetu, raziskavah in vsakdanjem življenju, saj omogoča pisanje besedil, ustvarjanje kode, generiranje slik in reševanje kompleksnih problemov. Možno je celo, da smo priča začetkom singularnosti – prelomnega trenutka, ko bo umetna inteligenca presegla človeško inteligenco in omogočila revolucijo na področju produktivnosti in inovacij, čeprav bo treba na sodbo o tem še počakati. Optimizem glede prihodnosti je utemeljen: če se bo razvoj nadaljeval s trenutnim tempom, si lahko predstavljamo svet, kjer bo umetna inteligenca povsem preoblikovala gospodarstvo, znanost in način življenja, pri čemer bo omogočila višjo kakovost življenja za vse.

Čeprav nekateri umetno inteligenco vidijo kot grožnjo, njen trenutni razmah resnejših težav še ni prinesel. Nadejamo se, da bo zadosten del raziskav usmerjen v varnost umetne inteligence, da bo tako ostalo. Z morebitnimi škodljivimi učinki umetne inteligence se spopadajo tudi regulatorji, za katere upamo, da bodo uspešno krmarili med tem ciljem in pretiranim zaviranjem razvoja.

Dostopnost velikih jezikovnih modelov, kot so GPT-ji, pomeni, da so naloge, ki zahtevajo razumevanje in generiranje naravnega jezika, lažje kot kadar koli prej. Mnogi raziskovalci verjamejo, da bo prihodnost programiranja prešla iz tradicionalnih jezikov, kot je Python, na velike jezikovne modele, kjer bo umetna inteligenca generirala kodo in rešitve po meri. Čeprav je razvoj teh modelov zahtevna naloga, ki presega zmožnosti večine organizacij, se ljudje navajamo na uporabo tega fenomenalnega orodja. Pričakujemo, da bo umetna inteligenca postala učinkovit in zanesljiv partner človeštva.

Že letos vidimo, da so konference v sklopu Informacijske družbe posvečene prav velikim jezikovnim modelom. V okviru Slovenske konference o umetni inteligenci organiziramo formalno debato dijakov – izkušenih debaterjev, ki se udeležujejo mednarodnih tekmovanj – o tem, kako bo umetna inteligenca oblikovala prihodnost in zakaj bi to lahko bila najboljša prihodnost doslej.

Matjaž Gams
Mitja Luštrek
Rok Piltaver
predsedniki Slovenske konference o umetni inteligenci

FOREWORD

Artificial intelligence is experiencing incredible and accelerated development. After three-quarters of a century since Alan Mathison Turing laid the foundations of computing and artificial intelligence, it is finally approaching not only human intelligence but also other key human traits such as creativity, emotional intelligence and consciousness. In many areas, artificial intelligence already surpasses the capabilities of most people and even experts. Large language models are achieving such results even in much less structured problems than was imaginable a few years ago, such as professional exams, and mathematics and programming tasks described in free text.

Generative artificial intelligence is already transforming the world. It has become an indispensable tool in the business world, research, and everyday life, enabling text writing, code generation, image creation, and solving complex problems. It is even possible that we are witnessing the beginnings of the singularity—the pivotal moment when artificial intelligence will surpass human intelligence and enable a revolution in productivity and innovation, although time will show whether this is actually the case. Optimism about the future is well-founded: if development continues at its current pace, we can imagine a world where artificial intelligence completely transforms the economy, science, and way of life, leading to a higher quality of life for all.

Although some see artificial intelligence as a threat, its current rapid progress has not yet led to serious problems. We hope that a sufficient part of the research will be directed towards AI safety so that this remains the case. Regulators are also addressing the potential harmful effects of artificial intelligence, and we hope they will successfully navigate between this goal and excessive hindering of development.

The accessibility of large language models, such as GPTs, means that tasks requiring the understanding and generation of natural language are easier than ever before. Many researchers believe that the future of programming will shift from traditional languages, like Python, to large language models, where artificial intelligence will generate custom code and solutions. Although developing these models is a challenging task beyond the capabilities of most organizations, people are getting accustomed to using this phenomenal tool. We expect artificial intelligence to become an effective and reliable partner for humanity.

Already this year, we are seeing conferences within the framework of the Information Society dedicated to large language models. As part of the Slovenian Conference on Artificial Intelligence, we are organizing a formal debate for high school students—experienced debaters who participate in international competitions—on how artificial intelligence will shape the future and why this might be the best future yet.

Matjaž Gams
Mitja Luštrek
Rok Piltaver
Slovenian Conference on Artificial Intelligence chairs

PROGRAMSKI ODBOR / PROGRAMME COMMITTEE

Mitja Luštrek

Matjaž Gams

Rok Piltaver

Cene Bavec

Marko Bohanec

Marko Bonač

Ivan Bratko

Bojan Cestnik

Aleš Dobnikar

Erik Dovgan

Bogdan Filipič

Borka Jerman Blažič

Marjan Krisper

Marjan Mernik

Biljana Mileva Boshkoska

Vladislav Rajkovič

Niko Schlamberger

Tomaž Seljak

Peter Stanovnik

Damjan Strnad

Miha Štajdohar

Vasja Vehovar

PandaChat-RAG: Towards the Benchmark for Slovenian RAG Applications

Taja Kuzman
Tanja Pavleska
{taja,tanja}@pc7.io
PC7, d.o.o.
Ljubljana, Slovenia
Jožef Stefan Institute
Ljubljana, Slovenia

Urban Rupnik
Primož Cigoj
{urban,primoz}@pc7.io
PC7, d.o.o.
Ljubljana, Slovenia

Abstract

Retrieval-augmented generation (RAG) is a recent method for enriching the large language models' text generation abilities with external knowledge through document retrieval. Due to its high usefulness for various applications, it already powers multiple products. However, despite the widespread adoption, there is a notable lack of evaluation benchmarks for RAG systems, particularly for less-resourced languages. This paper introduces the PandaChat-RAG – the first Slovenian RAG benchmark established on a newly developed test dataset. The test dataset is based on the semi-automatic extraction of authentic questions and answers from a genre-annotated web corpus. The methodology for the test dataset construction can be efficiently applied to any of the comparable corpora in numerous European languages. The test dataset is used to assess the RAG system's performance in retrieving relevant sources essential for providing accurate answers to the given questions. The evaluation involves comparing the performance of eight open- and closed-source embedding models, and investigating how the retrieval performance is influenced by factors such as the document chunk size and the number of retrieved sources. These findings contribute to establishing the guidelines for optimal RAG system configurations not only for Slovenian, but also for other languages.

Keywords

retrieval-augmented generation, RAG, embedding models, large language models, LLMs, benchmark, Slovenian

1 Introduction

The advent of large language models (LLMs) has introduced significant advancements in the field of natural language processing (NLP). Although LLMs have shown impressive capabilities in generating coherent text, they are prone to hallucinations [7, 16], i.e., providing false information. Furthermore, they are dependent on static and potentially outdated corpora [9]. Retrieval-augmented generation (RAG) is a method devised to address these challenges by augmenting LLMs with external information retrieved from a provided document collection. Connecting LLMs with a relevant database improves the factual accuracy and temporal relevance of the generated responses. Moreover, RAG contributes to the explainability of the generated answers by providing verifiable sources, which facilitates the evaluation of the system's accuracy [2]. These advantages have spurred quick adoption of RAG systems across various applications. For instance, PandaChat[1] leverages RAG to provide explainable responses with high accuracy in Slovenian and other languages, integrated in customer service bots and platforms that allow LLM-based retrieval of information from texts.

Although RAG benchmarking is a relatively recent endeavor, some initial frameworks have already emerged [3, 5, 7]. However, these benchmarks are only limited to English and Chinese, leaving a gap in the evaluation of RAG systems for other languages. To address this gap, we make the following contributions:

- We present the first benchmark for RAG systems for the Slovenian language. The benchmark is based on the newly developed PandaChat-RAG-sl test dataset[2], which comprises authentic questions, answers and source texts.
- We introduce a methodology for an efficient semi-automated development of RAG test datasets that is easily replicable for the languages included in the MaCoCu [1] and CLASSLA-web corpora collections [10], which include all South Slavic languages, Albanian, Catalan, Greek, Icelandic, Maltese, Ukrainian and Turkish.
- As the first step of RAG evaluation, we evaluate the retriever's performance in terms of its ability to provide relevant sources crucial to retrieve accurate answers to the posed questions. The evaluation encompasses comparison of performance of several open- and closed-source embedding models. Furthermore, we provide insights on the impact of the document chunk size and the number of retrieved sources, to identify optimal configurations of the indexing and retrieval components for robust and accurate retrieval.

The paper is organized as follows: in Section 2, we provide an introduction to the previous research concerning the evaluation of RAG systems; Section 3 introduces the PandaChat-RAG-sl dataset (Section 3.1) and the RAG system architecture (Section 3.2), which is evaluated in Section 4. Finally, in Section 5, we conclude the paper with a discussion of the main findings and suggestions for future work.

2 Related Work

Despite the recent introduction of the RAG architecture, several benchmarking initiatives have already emerged [3, 5, 7, 15]. However, since the RAG systems can be applied to various end tasks, the benchmarks focus on different aspects of these systems. Inter

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.538

[1] https://pandachat.ai/
[2] The PandaChat-RAG benchmark and its test dataset are openly available at https://github.com/TajaKuzman/pandachat-rag-benchmark.

alia, current benchmarks assess their performance in text citation [7], text continuation, question-answering with support of external knowledge, hallucination modification, and multi-document summarization [12].

The closest to our work is the evaluation of the RAG systems on the task of Attributable Question Answering [2]. This task involves providing a question as input to the system, which then generates both an answer and an attribution, indicating the source text on which the answer is based. The advantage of this task over the closed-book question-answering task is that it also measures the system's capability to provide the correct source.

The majority of RAG benchmarks assess RAG systems in English [3, 5, 7, 15] or Chinese [5, 12]. Consequently, the generalizability of their findings to other languages remains uncertain. Furthermore, a limitation of many benchmarks is their reliance on synthetic data generated by LLMs [5, 12, 15]. To avoid potential biases introduced by LLMs and to better represent the complexity and diversity of real-world language use, a more reliable evaluation would be based on non-synthetic test datasets. Despite focusing on a different task, recent research [6] has shown that resource-efficient development of non-synthetic and non-machine translated question-answering datasets is feasible by leveraging the availability of general web corpora and genre classifiers.

3 Methodology

3.1 PandaChat-RAG-sl Dataset

The PandaChat-RAG-sl dataset comprises questions, answers, and the corresponding source texts that encompass the answers. It was created through a semi-automated process involving the extraction of texts from the Slovenian web corpus CLASSLA-web.sl 1.0 [11], followed by a manual extraction of high-quality instances. Since the texts were automatically extracted from a general text collection, the dataset encompasses a diverse range of topics that were not predefined or decided upon.

The CLASSLA-web.sl 1.0 corpus is a collection of texts, collected from the web in 2021 and 2022 [10]. It was chosen due to its numerous advantages: 1) it has high-quality content, with the majority of texts meeting the criteria for publishable quality [17]; 2) it is one of the largest and most up-to-date collections of Slovenian texts, comprising approximately 4 million texts; 3) the texts are enriched with genre labels, facilitating genre-based text selection; and 4) it is developed in the same manner as 6 other CLASSLA-web corpora [10] and 7 additional MaCoCu web corpora in various European languages [1]. This enables easy expansion of the benchmark to other languages, including all South Slavic languages and various European languages, such as Albanian, Catalan, Greek, Icelandic, Ukrainian and Turkish.

The development of the PandaChat-RAG-sl dataset involves the following steps: 1) the genre-based selection of texts from the CLASSLA-web.sl corpus; 2) the extraction of texts that comprise paragraphs ending with a question (80,215 texts); 3) the extraction of questions and answers (paragraphs, following the question); 4) a manual review process to identify high-quality instances. In the genre-based selection phase, we extract texts labeled with genres that are most likely to contain objective questions and answers, that is, *Information/Explanation*, *Instruction* and *Legal*.

In its present iteration, the dataset consists of 206 instances derived from the first 1,800 extracted texts. It is important to note that this effort can easily be continued with further manual inspection of the extracted texts should there be a need to prepare a larger dataset.

Table 1: Statistics for the PandaChat-RAG-sl dataset.

	Number
Instances	206
Unique texts	160
Words (questions)	1,184
Words (texts w/o questions)	83,467
Total words (questions + texts)	84,651

Table 1 provides the statistical overview for the PandaChat-RAG-sl dataset. The dataset consists of 206 instances, that is, triplets of a question, an answer and a source text, derived from 160 texts. The total size of the dataset is 84,651 words, encompassing both the questions and the texts containing the answers.

3.2 RAG System

The RAG pipeline encompasses three main components: indexing, retrieval, and text generation. During the indexing phase, the user-provided text collection is transformed into a database of numerical vectors (embeddings) to facilitate document retrieval by the retriever. This process involves segmenting the documents into fixed-length chunks, which are then converted into embeddings using large language models. The choice of the embedding model and the chunk size are critical factors that can significantly impact the retrieval performance of the model. Selecting an appropriate embedding model is essential to ensure that the textual information is converted into a meaningful numerical representation for effective retrieval. Moreover, the chunk size, in terms of the number of tokens, plays a crucial role in determining the informativeness of the embeddings. Incorrect chunk sizes may lead to numerical vectors that lack important information necessary for connecting the question to the corresponding text chunk, thereby compromising retrieval accuracy [12].

When presented with a question, the retrieval component uses the semantic search (also known as dense retrieval) to retrieve the most relevant text chunks. The search is based on determining the smallest cosine distance between the chunk vectors and the question vector. Lastly, during the text generation phase, the retriever provides the large language model (LLM) with a selection of top retrieved sources. The LLM is prompted to provide a human-like answer to the provided question based on the retrieved text sources. The selection of an appropriate number of top retrieved sources is crucial in this phase: including more than just one retrieved source may enhance retrieval accuracy and address situations where the first retrieved source fails to encompass all relevant information, especially in the case when more texts cover the same subject matter. However, increasing the number of sources also leads to a longer prompt provided to the LLM, potentially increasing the costs of using the RAG system.

In this study, we assess the indexing and retrieval components, focusing on the impact of different embedding models, chunk sizes, and the number of retrieved sources on retrieval performance.

Embedding Models. The evaluation includes a range of multilingual open-source and closed-source models. The selection of open-source models is based on the Massive Text Embedding

Benchmark (MTEB) Leaderboard[3] [13]. Specifically, we choose medium-sized multilingual models with up to 600 million parameters that have demonstrated strong performance on Polish and Russian – Slavic languages that are linguistically related to Slovenian. The models used in the evaluation are:

- Closed-source embedding models provided by the OpenAI: an older model text-embedding-ada-002 (OpenAI-Ada) [8], and two recently published models: text-embedding-3-small (OpenAI-3-small), and text-embedding-3-large (OpenAI-3-large) [14].
- Open-source embedding models, available on the Hugging Face repository: BGE-M3 model [4], base-sized mGTE model (mGTE-base) [19], and small (mE5-small), base (mE5-base) and large sizes (mE5-large) of the Multilingual E5 model [18].

Chunk size. The impact of the chunk size on retrieval performance is assessed by varying chunk sizes of 128, 256, 512, and 1024 tokens, with a default chunk overlap of 20 tokens. In these experiments, the performance is evaluated based on the first retrieved source.

Number of retrieved sources. Previous work indicates that increasing the number of retrieved sources improves the retrieval accuracy [12]. In this study, we examine the retrieval accuracy of embedding models, with a chunk size set to 128 tokens, when the models retrieve 1 to 5 sources. In this scenario, if any of the multiple retrieved sources matches the correct source, the output is evaluated as being correct.

The retrieval capabilities of the RAG system are evaluated on the task of Attributed Question-Answering. The evaluation is based on accuracy, measured as the percentage of questions correctly matched with the relevant source.

The experiments are performed using the LlamaIndex library[4]. The chunk size is defined using the SentenceSplitter method in the indexing phase. Number of retrieved sources (*similarity top k*), the embedding model and the prompt for the LLM model are specified as parameters of the chat engine. The closed-source embedding models are used via the OpenAI API, while the experiments with the open-source models are conducted on a GPU machine.

4 Experiments and Results

In this section, we present the results of the experiments examining the impact of the chunk size, the number of retrieved sources, and the selection of the embedding model on the retrieval performance of the RAG system.

4.1 Chunk Size

Figure 1 shows the impact of the chunk size on the retrieval performance of the RAG systems that are based on different embedding models. The findings suggest that, with the exception of the OpenAI-Ada model, all systems demonstrate the best performance when the text chunk size is set to 128 tokens. Increasing the chunk size hinders the retrieval performance, which is consistent with previous research [12]. These results confirm that smaller chunk sizes enable the embedding models to capture finer details that are essential for retrieving the most relevant text for the given question.

[3]https://huggingface.co/spaces/mteb/leaderboard
[4]https://www.llamaindex.ai/

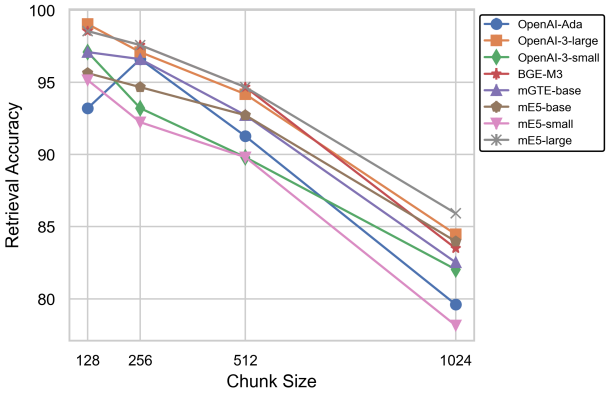


Figure 1: The impact of the chunk size on the retrieval performance.

4.2 Number of Retrieved Sources

Figure 2 shows the performance of the RAG systems when increasing the number of retrieved sources. The results demonstrate that increasing the number of retrieved sources initially improves the performance, however, after a certain threshold, the performance levels off.

Increasing the number of retrieved sources results in larger inputs to the LLM in the text generation component, incurring higher costs. Using more than two retrieved sources does not significantly improve results in most systems. What is more, with the top two retrieved sources, certain embedding models, namely, BGE-M3 and mE5-large, already reach perfect accuracy. Thus, our findings indicate that using more than the top two retrieved sources is unnecessary.

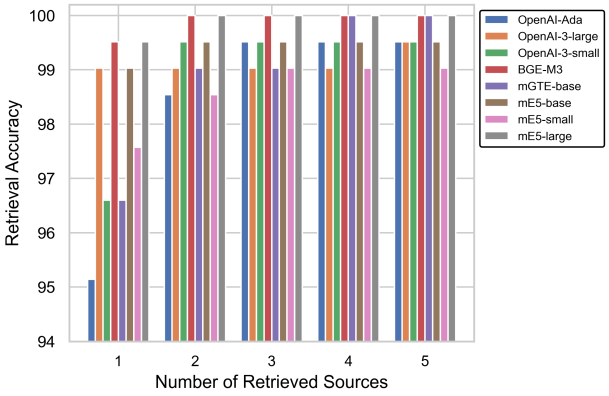


Figure 2: Impact of the number of retrieved sources on the retrieval performance.

4.3 Embedding Models

We provide the final comparison of the performance of systems that use different embedding models. We use the parameters that have shown to provide the best results in the previous experiments: the chunk size of 128 tokens and top two retrieved sources. As shown in Table 2, the retrieval systems that use the open-source BGE-M3 and mE5-large embedding models achieve the perfect retrieval score. They are closely followed by the closed-source OpenAI-3-small and the mE5-base models which achieve

Table 2: Performance comparison between the open-source and closed-source embedding models.

embedding model	speed (s)	retrieval accuracy
BGE-M3	0.58	100
mE5-large	0.58	100
OpenAI-3-small	0.69	99.51
mE5-base	0.29	99.51
OpenAI-3-large	1.19	99.03
mGTE-base	0.31	99.03
OpenAI-Ada	0.63	98.54
mE5-small	0.15	98.54

accuracy of 99.5%. While having slightly lower scores, all other retrieval systems still achieve high performance, ranging between 98.5% and 99% in accuracy.

Additionally, Table 2 provides the inference speed of the models measured in seconds per instance. If inference speed is a priority, the mE5-base model emerges as the optimal selection, as it yields high retrieval accuracy of 99.51% and is two times faster than the two best performing models. In cases where users are restricted to closed-source models due to the unavailability of GPU resources, the OpenAI-3-small model stands out as the most suitable option. Its inference speed is comparable to the OpenAI-Ada model, while it achieves a superior retrieval accuracy.

5 Conclusion and Future Work

In this paper, a novel test dataset was introduced to assess the performance of the RAG system on Slovenian language. A general methodology for efficient creating of non-synthetic RAG test datasets was established that can be extended to other languages. We evaluated the retrieval accuracy of the RAG system, examining the impact of the embedding models, the document chunk size, and the number of retrieved sources. The assessment of embedding models encompassed eight open-source and closed-source LLM models. It revealed that open-source models, specifically, BGE-M3 and mE5-large, reached perfect retrieval accuracy, demonstrating their suitability for RAG applications on Slovenian texts. Furthermore, the evaluation of the optimal chunk size and the number of retrieved sources showed that smaller chunk sizes yielded superior results. In contrast, increasing the number of retrieved sources enhanced results up to a certain threshold, beyond which the model performance plateaued. Certain models already achieved perfect accuracy when evaluated based on the top two retrieved sources.

While the novel test dataset can be used to evaluate all the components of the RAG system, in this paper, we focused on the evaluation of the indexing and retrieval components. In our future work, we will extend the evaluations to the text generation component with regard to fluency, correctness, and usefulness of the generated answers. Furthermore, we plan to expand the benchmark to encompass a wider range of languages. The plans include extending the dataset and evaluation to South Slavic languages and other European languages that are covered by comparable MaCoCu [1] and CLASSLA-web [10] corpora.

References

[1] Marta Bañón et al. 2022. MaCoCu: Massive collection and curation of monolingual and bilingual data: Focus on under-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, Ghent, Belgium, (June 2022), 303–304. https://aclanthology.org/2022.eamt-1.41.

[2] Bernd Bohnet et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

[3] Shuyang Cao and Lu Wang. 2024. Verifiable Generation with Subsentence-Level Fine-Grained Citations. *arXiv preprint arXiv:2406.06125*.

[4] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. (2024). arXiv: 2402.03216 [cs.CL].

[5] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 16. Vol. 38, 17754–17762.

[6] Anni Eskelinen, Amanda Myntti, Erik Henriksson, Sampo Pyysalo, and Veronika Laippala. 2024. Building Question-Answer Data Using Web Register Identification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors. ELRA and ICCL, Torino, Italia, (May 2024), 2595–2611. https://aclanthology.org/2024.lrec-main.234.

[7] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6465–6488.

[8] Ryan Greene, Ted Sanders, Lilian Weng, and Arvind Neelakantan. 2022. New and improved embedding model. https://openai.com/index/new-and-improved-embedding-model/. [Accessed 26-08-2024]. (2022).

[9] Angeliki Lazaridou et al. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34, 29348–29363.

[10] Nikola Ljubešić and Taja Kuzman. 2024. CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3271–3282.

[11] Nikola Ljubešić, Peter Rupnik, and Taja Kuzman. 2024. Slovenian web corpus CLASSLA-web.sl 1.0. In Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1882.

[12] Yuanjie Lyu et al. 2024. CRUD-RAG: A comprehensive Chinese benchmark for retrieval-augmented generation of large language models. *arXiv preprint arXiv:2401.17043*.

[13] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2014–2037.

[14] OpenAI. 2024. New embedding models and API updates. https://openai.com/index/new-embedding-models-and-api-updates/. [Accessed 26-08-2024]. (2024).

[15] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 338–354.

[16] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3784–3803.

[17] Rik van Noord, Taja Kuzman, Peter Rupnik, Nikola Ljubešić, Miquel Esplà-Gomis, Gema Ramírez-Sánchez, and Antonio Toral. 2024. Do Language Models Care about Text Quality? Evaluating Web-Crawled Corpora across 11 Languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 5221–5234.

[18] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint arXiv:2402.05672*.

[19] Xin Zhang et al. 2024. mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. (2024). https://arxiv.org/abs/2407.19669 arXiv: 2407.19669 [cs.CL].

Choosing Features for Stress Prediction with Machine Learning

Katja Bengeri
University of Ljubljana
Ljubljana, Slovenia
kb96968@student.uni-lj.si

Junoš Lukan*
Jožef Stefan Institute
Department of Intelligent Systems
Ljubljana, Slovenia
junos.lukan@ijs.si

Mitja Luštrek*
Jožef Stefan Institute
Department of Intelligent Systems
Ljubljana, Slovenia
mitja.lustrek@ijs.si

Abstract

Feature selection is a crucial step in building effective machine learning models, as it directly impacts model accuracy and interpretability. Driven by the aim of improving stress prediction models, this article evaluates multiple approaches for identifying the most relevant features. The study explores filter-based methods that assess feature importance through correlation analysis, alongside wrapper methods that iteratively optimize feature subsets. Additionally, techniques such as Boruta are analysed for their effectiveness in identifying all important features, while strategies for handling highly correlated variables are also considered. By conducting a comprehensive analysis of these approaches, we assess the role of feature selection in developing stress prediction models.

Keywords

Feature selection, Correlation matrix, Balanced accuracy score

1 Introduction

Machine learning models are increasingly being applied to predict stress, which is critical in various domains such as healthcare, workplace management, and wearable technology. However, one of the major challenges in developing reliable predictive models is identifying the most relevant features from extensive datasets, comprising physiological and behavioural information.

Feature selection plays a key role in addressing this challenge. By selecting only the most informative features, we can reduce noise, prevent overfitting, and enhance model accuracy. As we showed in previous work [8], even simple feature selection techniques can increase the F_1 score of predictive models. This paper builds upon this finding and explores several feature selection techniques, ranging from simple correlation-based methods to more sophisticated wrapper approaches.

The aim of this work is to assess how feature selection can enhance stress prediction models. By comparing different methods, we aim to identify the optimal strategies for feature selection in stress prediction which would lead to more reliable and more easily interpretable machine learning models.

2 Data collection

The data used in this work comes from the STRAW project [1], results of which have been previously presented at Information Society [6, 8]. The dataset includes the data of 56 participants, recruited from academic institutions in Belgium (29 participants) and Slovenia (26 participants). They answered questionnaires named Ecological Momentary Assessments (EMAs) roughly every 90 minutes, with smartphone sensor and usage data continuously collected by an Android application [7], while also wearing an Empatica E4 wristband recording physiological data. In 15 days of their participation, each participant responded to more than 96 EMA sessions, on average, which resulted in around 2200 labels.

3 Target and feature extraction

To fully leverage the potential of the data, we computed a comprehensive set of features. While some sensors only reported relatively rare events, such as phone calls, others had a high sampling frequency, such blood volume pulse which sampled data at 32 Hz. On the other hand, labels were only available every 90 min. Therefore, we preprocessed the data in several steps.

3.1 Target variable

While participants responded to various questionnaires, for this study, we selected their responses to Stress Appraisal Measurement [9] as the target variable. It was used to report stress levels on a scale from 0 to 4, so using it as is the prediction task can be approached as a regression problem.

However, many stress detection studies tend towards a discrete approach, treating stress predominantly as a classification task, often only working with a binary target variable. To convert this into a classification problem, we discretized the target variable into two distinct categories: "no stress", which included all responses with a value of 0, while all others were coded as "stress". With that, we ensured a balanced distribution of the target variable values.

3.2 Features

3.2.1 Data preprocessing. In our work, features were calculated on 30-minute intervals preceding each questionnaire session. From the wide variety of smartphone data and physiological measures, a total of 352 features were extracted and grouped into 22 categories, listed in Table 1. Using physiological data from Empatica wristband, we first calculated specialized physiological features on smaller windows (from 4 s to 120 s, depending on the sensor; see [4] for more details), which were then aggregated over 30 min windows by calculating simple statistical features: mean, median, standard deviation, minimum, and maximum. All of the categorical features were converted into a set of binary features using the one hot encoding technique and the missing values were replaced with the mode.

First, some preliminary data cleaning was performed by excluding one of the feature in pairs exhibiting a correlation coefficient of $|r| \geq 0.95$. Despite this, some of the remaining features still exhibited quite strong correlations as shown in Fig. 1. An interesting observation used in the later stages of feature selection was that high correlation, $|r| \geq 0.8$, was mostly observed

*Also with Jožef Stefan International Postgraduate School.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.991

Table 1: Feature categories with the number of features included in each category in parentheses

1. Empatica electrodermal activity (99)
2. Empatica inter-beat interval (50)
3. Empatica temperature (33)
4. Empatica accelerometer (23)
5. Empatica data yield (1)
6. Phone applications foreground (47)
7. Phone location (18)
8. Phone Bluetooth connections (18)
9. Phone calls (10)
10. Phone activity recognition (7)
11. Phone Wi-Fi connections (7)
12. Phone screen events (7)
13. Phone light (6)
14. Phone battery (5)
15. Phone speech (4)
16. Phone interactions (2)
17. Phone messages (2)
18. Phone data yield (1)
19. Baseline psychological features (7)
20. Language (2)
21. Gender (2)
22. Age (1)

between features of the same category. As an example, correlations between features related to phone application use are shown in Fig. 2.

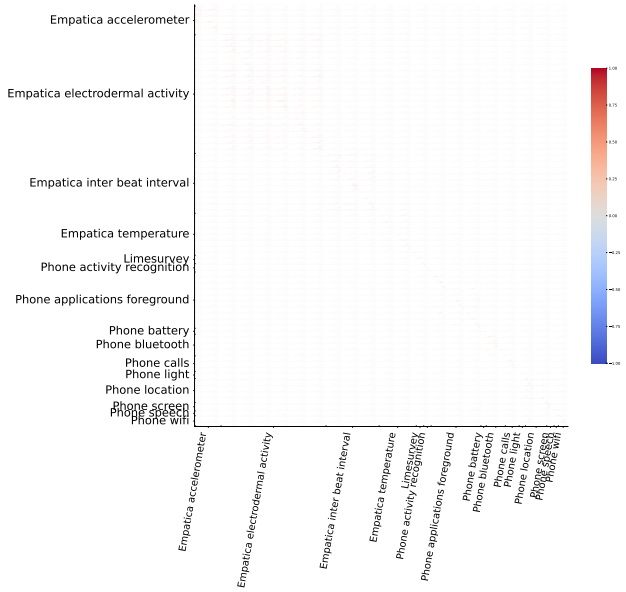


Figure 1: Correlation matrix of the initial feature set. Only feature categories with more than two features are labelled.

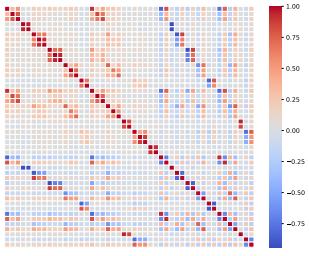


Figure 2: Correlation matrix of the feature set in the Phone applications foreground category.

4 Prediction models

4.1 Model performance and validation

To evaluate the performance of the models we used balanced accuracy score which is defined as the average of recall obtained on each class. When adjusted for random chance, it is calculated as

$$\textit{Balanced accuracy} = \frac{TP}{TP+FN} + \frac{TN}{TN+FP} - 1,$$

in the binary case, where TP is the number of true positives, TN is the number of true negatives, FN is the number of false negatives and FP is the number of false positives. This definition is equivalent to Youden's $\mathcal{J}$ [11], which assigns a 0 to a random classifier (indeed, a dummy classifier achieved a score of 0.0208 in our case), while a perfect classifier would achieve a score of 1.

To evaluate the stress detection models described in the following sections, we considered several ways of data partitioning. Since the variations in the results depending on the data split were significant, in order to achieve more consistent accuracy, we employed shuffled 5-fold cross-validation.

We also considered a leave-one-subject-out cross-validation technique. However, this method yielded poor results: using all available features, balanced accuracy was 0.05, while with the 5-fold cross validation it was 0.45. This suggested that the participants were quite different from each other, making it challenging to generalize predictions for a subject the model had not encountered.

4.2 Baseline model

Our initial approach for building a prediction model was to use all available features. This served as a baseline, which we aimed to improve through feature selection.

We evaluated various predictive models, as shown in Table 2, all as implemented in `scikit-learn` [10]. Among these, the Random Forest model yielded the best performance.

In this work, we aimed to find the best model for predicting stress and improve it using the optimal feature subset. Consequently, we used the Random Forest as the benchmark for comparing feature selection algorithms.

Table 2: Performance of different models for the classification problem. The mean over five folds, its standard error, and the maximum are shown.

Model	Mean	Max	SEM
Logistic Regression	0.077	0.151	0.025
Support Vector Machines	0.090	0.158	0.022
Gaussian Naive Bayes	0.061	0.122	0.020
Stochastic Gradient Descent	0.027	0.054	0.007
Random Forest	0.475	0.558	0.026
XGBoost	0.441	0.473	0.013

In Table 2, SEM represents the Standard Error of the Mean. It measures how far the sample mean of the data is likely to be from the true population mean.

4.3 Correlation-Based Feature Reduction

We began the feature selection process by eliminating highly correlated features. For each highly correlated pair, we removed the feature with the lower rank when sorted by mutual information, setting the correlation threshold at $|r| \geq 0.8$ to maintain a manageable number of features. This reduction left us with approximately 180 features out of the original 352 for model training and evaluation.

While selecting the optimal set of features for stress prediction, we aimed to retain all 22 different categories from Table 1, as

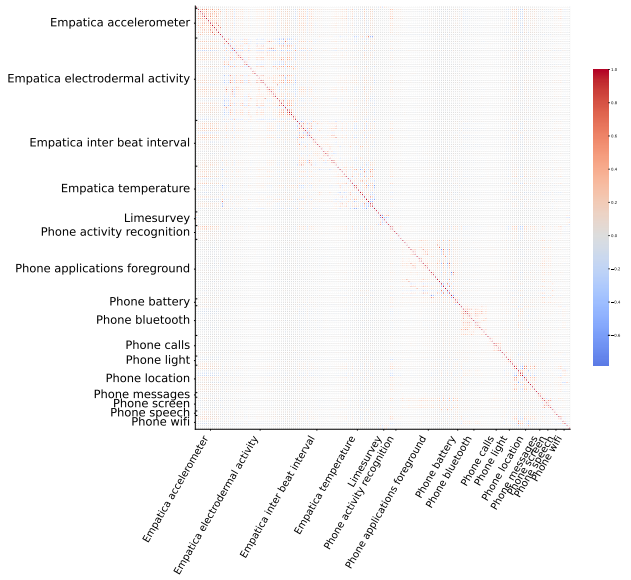


Figure 3: Correlation matrix of the feature set after correlation-based feature reduction. Only feature categories with more than two features are labelled.

each could provide unique information. Comparing Figs. 1 and 3, we were left with about half the number of features which were still moderately correlated.

4.4 Feature Selection using the mutual information scoring function

Before applying more complex feature selection algorithms, it was necessary to reduce computational complexity by further reducing our set of 180 features obtained through correlation-based reduction. Therefore, we used the `SelectKBest` method and the mutual information scoring function to retain the top 100 features. This resulted in features derived from 19 to 20 categories, as categories *language*, *gender*, and, in some cases, *Empatica accelerometer* were not deemed important for predicting stress.

Going forward, we will refer to the elimination of features within highly correlated pairs and the selection of the top 100 features using the mutual information scoring function as the preprocessing step.

4.5 Recursive Feature Elimination with Cross-Validation (RFECV)

One of the previously mentioned complex feature selection methods we employed was Recursive Feature Elimination with Cross-Validation (RFECV) [3]. The feature set we got after the preprocessing step was passed to the RFECV algorithm for thorough evaluation.

RFECV operates by initially fitting a model to the dataset and evaluating its performance through cross-validation. After the initial fit, RFECV ranks feature importance and iteratively removes the least important features based on the models feature importances attributes, which in the case of Random Forest are impurity-based feature importances. This process continues until there is no significant improvement in the model's performance. To ensure a reasonable duration for the feature selection process, we set the cross-validation in RFECV to 3 folds. The number of features selected varied across folds, ranging from 50 to 93 features.

4.6 Sequential Forward Selection

Another feature selection method we employed was Sequential Feature Selector (SFS), a wrapper-based technique [2]. SFS and RFECV differ in their approaches. SFS constructs models for each feature subset at every step, while RFECV builds a single model and evaluates feature importance scores. Consequently, SFS is more computationally expensive, as it must evaluate numerous feature combinations before identifying the optimal subset.

In the absence of specified parameters for number of features to select (`n_features_to_select`) and tolerance (`tol`), the method defaults to selecting half of the available features. The default configuration was used in our analysis, leading the SFS to select the top 50 features.

4.7 Boruta method

The final feature selection technique we employed was the Boruta method [5]. With the assistance of "shadow features", which are original features that have been randomly shuffled, the method identifies a subset of features that are relevant to the classification task at hand. In our case, shadow features were introduced into the feature subset obtained after the preprocessing step.

The updated dataset was trained using the Random Forest model for 100 iterations. In each iteration, all original features ranked higher in importance than the highest-ranked shadow feature were marked as relevant.

Ultimately, a binomial distribution is used to evaluate which features have enough confidence to be kept in the final selection. The number of features selected varied across folds, ranging from 47 to 55 features.

5 Results

In Table 3, the final scores for a Random Forest model built on various feature subsets, as derived from the methods described above, are presented. The data was split using shuffled 5-fold cross-validation, to ensure that the results were not overly dependent on a data split.

Table 3: Adjusted balanced accuracy scores of a Random Forest model, trained on the different feature sets. Last column represents a number of features selected.

Feature set	Mean	Max	SEM	N
All available features	0.464	0.498	0.011	352
Correlation-based reduction	0.483	0.507	0.007	~180
Correlation-based, 100 best	0.486	0.498	0.006	100
Preprocessing, RFECV	0.471	0.511	0.012	50 to 93
Preprocessing, SFS	0.483	0.520	0.017	50
Preprocessing, Boruta	0.481	0.545	0.020	47 to 55
RFECV only	0.465	0.494	0.020	16 to 89
SFS only	0.426	0.468	0.017	30
Boruta only	0.456	0.509	0.015	~75

From Table 3, we can see that the most significant improvement in accuracy came after removing the highly correlated features, with the average adjusted balanced accuracy score rising from 0.46 to 0.48. Best mean accuracy was achieved after the preprocessing step, with only a minor improvement from 0.483 to 0.486.

After eliminating highly correlated features, wrapper methods did not significantly improve the accuracy on average (rows 3 to 6 in Table 3). The Boruta method performed best among the three, with the highest overall maximum accuracy in a single fold. These results led us to investigate whether the wrapper feature selection method alone could manage correlated features without their prior removal and to evaluate the impact of the correlation threshold.

We employed the RFECV, SFS, and Boruta method on the entire feature set of 352 features without applying the preprocessing step. For SFS, only 30 features were selected due to its computational complexity. As shown in the last three rows of Table 3, none of the methods alone were able to improve the result achieved with correlation removal. Highly correlated features were left in the final feature set: for example, we identified three pairs of features with a correlation coefficient exceeding $|r| \geq 0.8$ using SFS alone. Poor results could be attributed either to the importance of the correlation removal step or to the feature subset being too small in the case of the SFS.

5.1 Selecting the best correlation threshold

As previously mentioned, the biggest improvement in score came from removing the feature inside the highly correlated pair. Therefore, we have also experimented with different correlation cut-off values to determine the best threshold.

The highest score was achieved with a correlation threshold of $|r| \geq 0.75$ (Table 4). Considering the impact of cross-validation splits and the relatively minor variance in scores, it appears that our initial threshold of $|r| \geq 0.8$ was also quite effective.

Table 4: Adjusted balanced accuracy scores of a Random Forest model trained on a feature subset excluding features above the correlation threshold. The number of features left after correlation-based feature selection differed over validation folds and its range is shown in the final column.

Threshold	Mean	Max	SEM	N
0.55	0.462	0.506	0.018	28 to 33
0.60	0.467	0.493	0.009	39 to 41
0.65	0.474	0.498	0.008	47 to 50
0.70	0.460	0.501	0.017	61 to 65
0.75	0.498	0.526	0.012	74 to 80
0.80	0.470	0.543	0.022	101 to 107

6 Conclusions

This paper examined different feature selection algorithms to find the most effective subset for stress prediction. The model using the feature subset after correlation removal achieved the highest adjusted balanced accuracy score of 0.483.

Alternative feature selection approaches, including the wrapper methods SFS and RFECV, as well as the Boruta method, applied to the preprocessed feature subset, did not lead to further optimization of the feature subset in terms of model performance. Additionally, applying these methods to the entire set of features did not achieve accuracy levels as high as those obtained after the correlation-based reduction. In the case of SFS, this may be attributed to its selection of only 30 features.

Therefore, our results underscore the critical role of the correlation-based reduction step. In contrast, when this step was omitted wrapper methods alone were unable to effectively perform correlation-based feature reduction. We can therefore conclude that simply relying on feature selection methods, however sophisticated, is not as effective as also considering relationships between features.

It should be noted that the improvements in balanced accuracy are low in all cases. This indicates that results cannot be easily generalized and correlation-based feature selection should not be seen as sufficient in general. Instead, we can speculate that no single feature selection method is the best one and that several should be considered. We should also note that the Pearson correlation coefficient that we used in this work only considers linear relationships between features. Other methods can select features even if they are related in a different way.

References

[1] Larissa Bolliger, Junoš Lukan, Mitja Luštrek, Dirk De Bacquer, and Els Clays. 2020. Protocol of the STRess at Work (STRAW) project: how to disentangle day-to-day occupational stress among academics based on EMA, physiological data, and smartphone sensor and usage data. *International Journal of Environmental Research and Public Health*, 17, 23, (Nov. 2020), 8835. DOI: 10.3390/ijerph17238835.

[2] Francesc J. Ferri, Pavel Pudil, Mohamad Hatef, and Josef V. Kittler. 1994. Comparative study of techniques for large-scale feature selection. *Machine Intelligence and Pattern Recognition*, 16, 403–413. DOI: 10.1016/b978-0-444-81892-8.50040-7.

[3] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using Support Vector Machines. *Machine Learning*, 46, 1/3, 389–422. DOI: 10.1023/a:1012487302797.

[4] Vito Janko, Matjaž Boštic, Junoš Lukan, and Gašper Slapničar. 2021. Library for feature calculation in the context-recognition domain. In *Proceedings of the 24th International Multiconference Information Society – IS 2021*. Slovenian Conference on Artificial Intelligence (Ljubljana, Slovenia, Oct. 4–8, 2021). Vol. A, 23–26.

[5] Miron B. Kursa and Witold R. Rudnicki. 2010. Feature selection with the Boruta package. *Journal of Statistical Software*, 36, 11, 1–13. DOI: 10.18637/jss.v036.i11.

[6] Junoš Lukan, Larissa Bolliger, Els Clays, Primož Šiško, and Mitja Luštrek. 2022. Assessing sources of variability of hierarchical data in a repeated-measures diary study of stress. In *Proceedings of the 25th International Multiconference Information Society – IS 2022*. Pervasive Health and Smart Sensing (Ljubljana, Slovenia, Oct. 10–14, 2022). Vol. A, 31–34.

[7] Junoš Lukan, Marko Katrašnik, Larissa Bolliger, Els Clays, and Mitja Luštrek. 2020. STRAW application for collecting context data and ecological momentary assessment. In *Proceedings of the 23rd International Multiconference Information Society – IS 2020*. Slovenian Conference on Artificial Intelligence (Ljubljana, Slovenia, Oct. 5–9, 2020). Vol. A, 63–67.

[8] Marcel Franse Martinšek, Junoš Lukan, Larissa Bolliger, Els Clays, Primož Šiško, and Mitja Luštrek. 2023. Social interaction prediction from smartphone sensor data. In *Proceedings of the 26th International Multiconference Information Society – IS 2023*. Slovenian Conference on Artificial Intelligence (Ljubljana, Slovenia, Oct. 9–13, 2023). Vol. A, 11–14.

[9] Edward J. Peacock and Paul T. P. Wong. 1990. The stress appraisal measure (SAM). A multidimensional approach to cognitive appraisal. *Stress Medicine*, 6, 3, (July 1990), 227–236. DOI: 10.1002/smi.2460060308.

[10] F. Pedregosa et al. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[11] Charles Sanders Peirce. 1884. The numerical measure of the success of predictions. *Science*, ns-4, 93, 453–454. DOI: 10.1126/science.ns-4.93.453.b.

Predictive Modeling of Football Results in the WWIN League of Bosnia and Herzegovina

Ervin Vladić
International Burch University
Sarajevo, Bosnia and Herzegovina
ervin.vladic@stu.ibu.edu.ba

Dželila Mehanović
International Burch University
Sarajevo, Bosnia and Herzegovina
dzelila.mehanovic@ibu.edu.ba

Elma Avdić
International Burch University
Sarajevo, Bosnia and Herzegovina
elma.avdic@ibu.edu.ba

Abstract

Predictive modeling in football has emerged as a valuable tool for enhancing decision-making in sports management. This study applies machine learning techniques to predict football match outcomes in the WWIN League of Bosnia and Herzegovina. The aim is to evaluate the effectiveness of various models, including Support Vector Machines (SVM), Logistic Regression, Random Forest, Gradient Boosting, and k-Nearest Neighbors (kNN), in accurately predicting match results based on key features such as shots on target, possession percentage, and home/away status. By (1) gathering and analyzing match data from three seasons, (2) comparing the performance of machine learning models, and (3) drawing conclusions on key performance factors, we demonstrate that SVM achieves the highest accuracy at 83%, outperforming other models. These insights contribute to football management, allowing for data-driven strategic planning and performance optimization. Future research will integrate additional factors such as player injuries and weather conditions to improve the predictive models further.

Keywords

Football match prediction, machine learning, WWIN league, Support Vector Machines, Random Forest

1 Introduction

Accurate predictions of match outcomes can inform a wide range of decisions, from tactical adjustments to player acquisitions, and can improve engagement for fans and stakeholders. While predictive modeling has been extensively applied to top-tier football leagues like the English Premier League, there is limited research on regional leagues such as the WWIN League of Bosnia and Herzegovina. The specificity of the country that is Bosnia and Herzegovina and the WWIN League, which has not been researched in the sphere of sports research, provides context for this step.

The WWIN League of Bosnia and Herzegovina was established in the year 2000 and the same year the WWIN was formed by the merging of three leagues, it became a league covering the entire territory of Bosnia and Herzegovina. Originally, the league consisted of 16 clubs, and, from the 2016-2017 season, the league contains 12 clubs which makes the level of the league higher [25]. The winner is the team that has the most points by the completion of thirty-three rounds; this position will grant a team a place in the UEFA Champions League qualifications [10]; the remaining two teams and the winner of the cup will compete for a place in the UEFA Conference League. Since the founding of the WWIN League of Bosnia and Herzegovina, team with the highest number of titles was HŠK Zrinjski from Mostar who emerged as the winner eight times, followed by Sarajevo which won four times, Zeljeznicar and Borac both won three times, Siroki Brijeg won two times and Leotar and Modrica both won once [12]. Depending on which entity association they belong to, the teams that occupy the last two places in the league at the end of the season are relegated to the league below, with two teams from the First League of the Federation of BiH and the First League of the RS being promoted in their stead. To elevate football in our country to the highest level, we must support in-depth analyses of matches and the factors influencing their outcomes. This will enable coaches to fine-tune strategies for future games, help commentators provide more insightful commentary, and allow fans to develop a deeper understanding and get more pleasure from the match.

The study aims to evaluate the performance of various ML models, including Support Vector Machines (SVM), Logistic Regression, Random Forest, Gradient Boosting, and k-Nearest Neighbors (kNN), in predicting match results. By examining key features such as shots on target, possession percentage, and home/away status, we conduct an analysis based on match data from three seasons of the WWIN League, encompassing 400 matches and key performance metrics.

The remainder of the paper is structured as follows: Section II provides an overview of related work in football ML-based prediction. Section III describes the methodology, including the dataset and models used. Section IV presents the results and analysis of models performance, with a discussion on the practical implications of the findings for football management. Finally, Section V concludes the paper and outlines directions for future research.

2 Literature Review

The prediction of the results of football matches has been recently studied extensively because of its relation to betting and decision-making in sports. Studies examining the employment of ML methods are primarily focused on large European leagues, where extensive and highly detailed data is available. The application of these techniques to regional football leagues, such as the WWIN League of B&H, remains underexplored.

Rodrigues and Pinto [15] used a variety of ML methods, including Naive Bayes, K-nearest neighbors, Random Forest, and SVM, to predict the match outcomes based on previous match data and player attributes. Their studies revealed excellent results in terms of soccer betting profit margins, with the Random Forest approach obtaining a success rate of 65.26% and a profit margin of 26.74%. Rahman [13] dedicated his work to employing deep learning frameworks especially Deep Neural Networks (DNNs) for football match outcome prediction, particularly during FIFA World Cup 2018. The study classified match outcomes

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.1642

with 63.3% accuracy with DNN architectures with LSTM or GRU cells. Baboota and Kaur [3] used machine learning approaches to predict English Premier League match results. The models compared included Support Vector Machines, Random Forest; and Gradient Boosting. From their study , they ascertained that Gradient Boosting outperformed other models in accuracy and overall predictiveness. Authors in [16] used machine learning techniques, notably SVM and Random Forest Classifier, to predict English Premier League (EPL) football match results. They got 54.3% accuracy with SVM and 49.8% with Random Forest after evaluating data from 2013/2014 to 2018/2019 seasons. Another study [8] employed a few machine learning algorithms to predict matches of the English Premier League season 2017-2018. Models including Linear Regression, SVM, Logistic Regression, Random Forest, and Multinomial Naïve Bayes classier show that the K-nearest neighbors give the best accurate predictions.

In summary, while existing studies have demonstrated the effectiveness of machine learning in football matches prediction, there remains a gap in the application of these techniques to regional leagues like the WWIN League, due to the availability and quality of data. The characteristics of these leagues, such as smaller datasets and potentially different factors influencing match outcomes, require a tailored approach. In lesser-known football leagues models might perform differently due to variations in competitive structures and gameplay strategies, as well. The study of Munđar and Šimić [11] in which they developed a simulation model using the Poisson distribution to predict the seasonal rankings of teams in the Croatian First Football League, highlighted the predictive power of statistical models and demonstrated the significance of home advantage in determining match outcomes, which is also an important factor in the WWIN League.

3 Materials and Methods

In this section, we describe the study conducted, detailing the data collection and feature selection processes, the machine learning models applied, the evaluation metrics used to assess model performance, and the approach taken to analyze key features influencing match outcomes. As a result of providing numerous procedures that are declared in this section, we represent the graphical illustration of our methodology. The steps involved in predicting the outcomes of the WWIN League of Bosnia and Herzegovina, including data collection, preprocessing, model development, and algorithm evaluation.

3.1 Dataset

The authors created the dataset for this study by consolidating information from rezultati.com [14], 1XBET [1], and Sofascore [24]. The unique dataset represents the seasons 2021/2022, 2022/2023, and 2023/2024 of WWIN League of Bosnia and Herzegovina. The platforms provide a wide range of football match data so it is easy to find important information for examination. The dataset includes key match facts as date, day of the week, time, home team, away team, final as well as half-time goals scored in the game, referee details, shots taken at goal as well as corner kicks resulting from these attempts on target, bookings made during play by both teams and other relevant performance indicators.

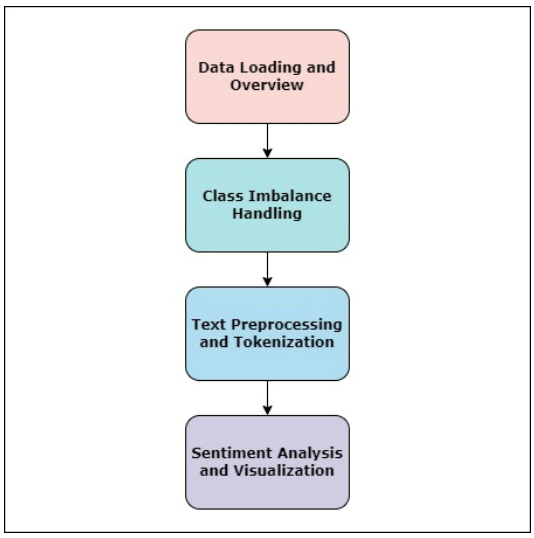


Figure 1: Workflow diagram

Table 1: Class Distribution

Match Type	Count
Home Win	301
Away Win	142
Draw	151

The table sums up a type of match result in terms of its frequency in the dataset.

In the recorded 594 matches, 301 ended in home team victories, 142 in away team victories and 151 were tied. The following pie chart describes the percentage distribution of the match outcome. Curiously, home wins are in the majority, comprising 50.7% of all

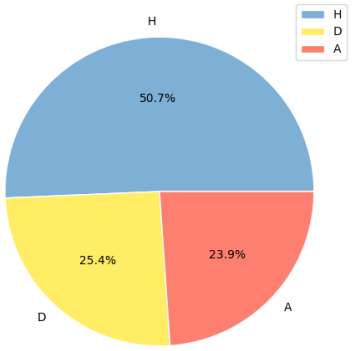


Figure 2: Class distribution of the dataset

matches. However, away victories contribute to approximately 23.9% of all recorded match results, while 25.4% contribute to draw results. The Fig.2 depicts the frequency of each of the match outcomes.

3.2 Machine Learning Prediction

In football, the concept of machine learning prediction entails developing models to forecast match outcomes based on the teams'

and players' histories and other attributes [5]. These models employ such methods as regression analysis, classification, and neural nets to determine the results given the data fed as the input.

3.2.1 Models initialisation, preprocessing, training and testing. While implementing Logistic Regression, we have set the max_iter =1000 and random_state = 42. Again, with the same classifier, the kernel argument was assigned a linear value while the random_state was set to 42 to keep the results predictable. Gaussian Naive Bayes was employed with no modification of its settings because of the model's simplistic nature. For Random Forest, we used the default parameters since the algorithm is capable of changing the setting on its own based on the complexity of given data. We initiated the Gradient Boosting with the default parameters so that the gradients could easily learn and an ensemble could be formed. Last but not least, we left all the parameters of k-Nearest Neighbors (kNN) for default value because the algorithm can find the optimal number of neighbors appropriate for the distribution of the data.

Following that, we proceed with the process of dividing this gathered data into two sets: the training and the testing ones. We split the data into training, where 70% of the data was allocated and the testing data where 30% was allocated.

Subsequently, the phase of model preprocessing is created for which it is essential to filter data effectively to ensure proper model training. In the case of feature transformation, we used scikit-learn's ColumnTransformer [17] to empower the numeric features normalization via the StandardScaler [23] while transforming the categorical variables into the binary format by the use of the OneHotEncoder [18]. This technique pays a lot of attention to ensuring that feature types are standard as well as harmonious. This method ensures consistency by creating a pipeline where preprocessing processes and the model are joined in the same line of work. This means that there is always uniformity in the training and the testing of the model, hence a manageable variability. Assuming the pipeline has been defined and is ready to proceed, we proceed to the next step of model training.

3.2.2 Models in Detail. In this study, many supervised learning classifier techniques that have proven valuable in the sports area for predictive purposes are employed. Logistic Regression is a statistical technique that predicts the probability of a binary classification, using a sigmoid function to map outputs to a [0,1] probability space. Coefficients indicate the strength and direction of relationships between variables, with positive values increasing the likelihood of an event and negative values decreasing it [9].

Random Forest extends the bagging method by generating multiple decision trees using randomly selected data samples. Each tree operates independently, and the final prediction is the average result across all trees, reducing overfitting and improving accuracy in classification tasks [4].

SVM aims to find the best hyperplane to separate data points by class, maximizing the margin between them. It handles nonlinear boundaries by transforming the input data into a higher-dimensional space [2].

Naïve Bayes, based on Bayes' theorem, assumes feature independence, making it fast and easy to implement, especially in applications like spam detection and text classification. Despite the simplicity of this assumption, it performs well in practice [26].

Gradient Boosting combines multiple weak learners (typically decision trees) to create a stronger predictive model, improving accuracy by focusing on correcting errors from previous models [6].

k-Nearest Neighbors (kNN) is an instance-based learning method that classifies data by identifying the majority label among the k closest points. Though simple, it can be computationally expensive as it requires storing all training data and performing real-time comparisons [7].

3.2.3 Evaluation Metrics. Last but not the least, the trained models are assessed by metrics such as accuracy of the models [19], precision of the models [21], the recall of the models [22], and F1-score value of the models [20]. This evaluation enables one to analyze how well each of the models is likely to perform in terms of match outcome prediction.

4 Results and Discussion

In this study, we employed six different classifiers to predict football match outcomes and conducted a comparative analysis of their performance. The effectiveness of each classifier was evaluated based on its accuracy, providing a clear comparison of their predictive capabilities across the dataset.

4.1 Model Performance

Among the classifiers employed, SVM predicted the most accurate results at 83% This model performed almost well, with balanced precision and recall across all three classes (A, D, and H), showing that it can predict match outcomes. In comparison, Random Forest achieved a lower accuracy of 65%, with especially evident deficits in precision and recall for class 'D'. Logistic Regression performed worse than Support Vector Machines, with accuracy of 77%. Despite its simplicity and computational efficiency, Gaussian Naive Bayes had the lowest accuracy of any classifier tested, at 39%. This model struggled to predict class 'D', with low accuracy and recall scores. Random Forest, an established ensemble learning approach, performed not so good, with an accuracy of 54%. This model has generally balanced accuracy and recall across all classes, making it an acceptable alternative for predicting match results. Gradient Boosting, another ensemble learning technique, has a little higher accuracy than Random Forest at 64%. While Gradient Boosting is recognized for its ability to manage complicated connections, it produced poorer recall ratings, especially for class 'D'. Lastly, k-Nearest Neighbors (kNN) resulted in 51% accuracy, showing that the classifier was relatively poor, they had relatively fair precision and recall with all the classes.

For making the match predictions, we employed the following classification models – Logistic Regression, Support Vector Machine, Gaussian Naive Bayes, Random Forest, Gradient Boost and k-Nearest Neighbors. We obtained the results varying from 39% to 83%, in which Support Vector Machines were the most effective. Our findings are partially consistent with prior research because classifiers like Support Vector Machines, Logistic Regression, and Random Forest have manifested robustness in predicting the match outcome across datasets. Nevertheless, the results are not in conformity with some emerging works, particularly concerning the efficacy of Gaussian Naive Bayes, which performed poorly in our study in contrast to other research results. It should be noted that results may vary significantly between different studies depending on the quality, the quantity, and the nature of the data that had been used for creating the models of Gaussian Naive Bayes.

Model	Accuracy	Precision	Recall	F1 score
Logistic Regression	77%	75%	74%	74%
SVM	83%	86%	83%	84%
Gaussian NB	39%	47%	42%	36%
Random Forest	54%	43%	46%	43%
Gradient Boosting	64%	64%	59%	60%
kNN	51%	49%	49%	49%

Table 2: Model Performances

The Table 2 shows how accurate various machine learning models are in predicting WWIN League of Bosnia and Herzegovina match outcomes.

4.1.1 Key factors influencing match outcomes. While this study does not perform formal feature analysis, the observed performance trends allow us to draw conclusions about the key factors influencing match outcomes. In line with prior research, home advantage emerged as a critical factor, with teams winning at home in over 50% of cases (Table 1) which reinforces the psychological and tactical advantages that come with playing on familiar ground.

Offensive metrics, particularly shots on target, also revealed themselves as strong predictors of success. Teams that generated more attempts on goal were significantly more likely to win, reinforcing the widely accepted view that aggressive, forward-driven play translates directly into better results. This trend mirrors observations from other football leagues, where offensive intensity is often directly correlated with victory.

4.1.2 Limitations and future work. Despite the promising results, this study has several limitations. First, the dataset used does not account for external factors such as player injuries, weather conditions, or team morale, all of which can influence match outcomes. Future research should incorporate these factors to improve the accuracy of predictions. Second, while SVM performed well in this context, more advanced models such as deep learning could potentially offer even better predictive performance, particularly when dealing with larger datasets.

Future work will explore the integration of additional domain-specific features, such as player statistics, team form, and environmental conditions, to further refine the predictive models. We will also experiment with more complex algorithms, such as neural networks, to capture the intricate relationships between features that may be missed by traditional machine learning models.

5 Conclusion

This study demonstrates that machine learning, particularly SVM, effectively predicts football match outcomes in the WWIN League of Bosnia and Herzegovina. Support Vector Machine has been found to be the highest accurate classifier with 83% of accuracy rate on match result prediction. SVM has moderate accuracy and recall with all three outcome classes: Home Win, Away Win, and Draw, indicating football prediction applicability. However, it has also revealed that other classifiers' performances are varying with Logistic Regression producing 77% of accuracy and Gaussian Naïve Bayes a poor 39% accuracy. Both Random Forest and Gradient Boosting, which are ensemble learning algorithms, have similar levels of accuracy; 54% and 64% respectively. While further refinement of the models is needed, the current findings establish a strong foundation for data-driven decision-making in football management. Future work should incorporate additional factors such as player injuries and weather conditions to enhance predictive accuracy.

References

[1] 1XBET. 2007–2024. 1xbet. Retrieved May 26, 2024, from https://1xlite-579542.top/en?tag=s_245231m_5435c_. (2007–2024).

[2] Mariette Awad and Rahul Khanna. 2015. Support vector machines for classification. In *Efficient Learning Machines*. Rahul Khanna, editor. Apress, 39–66. DOI: 10.1007/978-1-4302-5990-9_3.

[3] Rahul Baboota and Harleen Kaur. 2019. Predictive analysis and modeling football results using a machine learning approach for the english premier league. *International Journal of Forecasting*, 35, 2, 741–755. DOI: 10.1016/j.ijforecast.2018.01.003.

[4] Leo Breiman. 2001. Random forests. *Machine Learning*, 45, 5–32. DOI: 10.1023/A:1010933404324.

[5] Rory P. Bunker and Fadi Thabtah. 2019. A machine learning framework for sport result prediction. *Applied Computing and Informatics*, 15, 1, 27–33.

[6] Stefanos Fafalios, Pavlos Charonyktakis, and Ioannis Tsamardinos. 2020. *Gradient Boosting Trees*. Gnosis Data Analysis PC, (Apr. 2020).

[7] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. 2003. Knn model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003*. Springer Berlin Heidelberg, Catania, Sicily, Italy, (Nov. 2003), 986–996.

[8] Ishan Jawade, Rushikesh Jadhav, Mark Joseph Vaz, and Vaishnavi Yamgekar. 2021. Predicting football match results using machine learning. *International Research Journal of Engineering and Technology (IRJET)*, 8, 7, 177. https://www.irjet.net.

[9] Daniel Jurafsky and James H. Martin. 2023. *Logistic Regression*. Stanford University, 5. https://web.stanford.edu/~jurafsky/slp3/5.pdf.

[10] Haris Kruskic. 2019. Uefa champions league explained: how the tournament works. Bleacher Report. Retrieved from https://bleacherreport.com/articles/2819840-uefa-champions-league-explained-how-the-tournament-works. (2019).

[11] Dušan Munđar and Diana Šimić. 2016. Roatian first football league: teams' performance in the championship. *roatian Review of Economic, Business and Social Statistics 2*, 2, 1, 15–23. https://hrcak.srce.hr/file/245359.

[12] Prva Liga BiH. 2022. Osvajači trofeja. Retrieved from https://plbih.ba/osvajaci-trofeja/. (2022).

[13] Ashiqur Rahman. 2020. A deep learning framework for football match prediction. *SN Applied Sciences*, 2, 2, 165. DOI: 10.1007/s42452-019-1821-5.

[14] 2006–2024. Rezultati. Retrieved May 26, 2024, from https://www.rezultati.com/. (2006–2024).

[15] Fátima Rodrigues and Ângelo Pinto. 2022. Prediction of football match results with machine learning. *Procedia Computer Science*, 204, 463–470. DOI: 10.1016/j.procs.2022.08.057.

[16] Sayed Muhammad Yonus Saiedy, Muhammad Aslam HemmatQachmas, and Dr. Amanullah Faqiri. 2020. Predicting epl football matches results using machine learning algorithms. *International Journal of Engineering Applied Sciences and Technology*, 5, 3, 83–91. http://www.ijeast.com.

[17] scikit-learn. 2024. Columntransformer. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.compose.ColumnTransformer.html. (2024).

[18] scikit-learn. 2024. Onehotencoder. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html. (2024).

[19] scikit-learn. 2024. Sklearn.metrics.accuracy_score. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html. (2024).

[20] scikit-learn. 2024. Sklearn.metrics.f1_score. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html. (2024).

[21] scikit-learn. 2024. Sklearn.metrics.precision_score. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html. (2024).

[22] scikit-learn. 2024. Sklearn.metrics.recall_score. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html. (2024).

[23] scikit-learn developers. 2024. Sklearn.preprocessing.standardscaler. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. (2024).

[24] Sofascore. 2024. Sofascore. Retrieved May 26, 2024, from https://www.sofascore.com/. (2024).

[25] SportMonks. 2022. Premier league api bosnia. Retrieved from https://www.sportmonks.com/football-api/premier-league-api-bosnia/. (2022).

[26] Geoffrey I. Webb. 2016. Naïve bayes. In *Encyclopedia of Machine Learning and Data Mining*. Claude Sammut and Geoffrey I. Webb, editors. (Jan. 2016), 1–2. DOI: 10.1007/978-1-4899-7502-7_581-1.

Sarcasm Detection in a Less-Resourced Language

Lazar Đoković
lazardjokoviclaki02@gmail.com
University of Ljubljana, Faculty of Computer and Information Science
Ljubljana, Slovenia

Marko Robnik-Šikonja
marko.robnik@fri.uni-lj.si
University of Ljubljana, Faculty of Computer and Information Science
Ljubljana, Slovenia

Abstract

The sarcasm detection task in natural language processing tries to classify whether an utterance is sarcastic or not. It is related to sentiment analysis since it often inverts surface sentiment. Because sarcastic sentences are highly dependent on context, and they are often accompanied by various non-verbal cues, the task is challenging. Most of related work focuses on high-resourced languages like English. To build a sarcasm detection dataset for a less-resourced language, such as Slovenian, we leverage two modern techniques: a machine translation specific medium-size transformer model, and a very large generative language model. We explore the viability of translated datasets and how the size of a pretrained transformer affects its ability to detect sarcasm. We train ensembles of detection models and evaluate models' performance. The results show that larger models generally outperform smaller ones and that ensembling can slightly improve sarcasm detection performance. Our best ensemble approach achieves an F_1-score of 0.765 which is close to annotators' agreement in the source language.

Keywords

natural language processing, large language models, sarcasm detection, neural machine translation, BERT model, GPT model, LLaMa model, ensembles

1 Introduction

Sentiment analysis is a popular task in natural language processing (NLP), concerned with the extraction of underlying attitudes and opinions, usually categorized as "positive", "negative", and "neutral". Detection of sentiment is challenging if the utterances are ironic or sarcastic. *Sarcasm* is a form of verbal irony that transforms the surface polarity of an apparently positive or negative utterance/statement into its opposite [6]. Sarcasm is frequent in our day-to-day communication, especially on social media [5]. This poses a significant problem for sentiment analysis tools since sarcasm polarity switches create ambiguity in meaning. Sarcasm is highly dependent on its context. For example, the sentence "*I just love hot weather*" could be interpreted as sarcastic, depending on the situation, e.g., during summer heat waves.

Historical developments of sarcasm detection are surveyed by Joshi et al. [3], while recent developments are covered by Moores and Mago [5]. The problem of automatic sarcasm detection in text is most commonly formulated as a classification task. Unfortunately, sarcasm detection is affected by the lack of large-scale, noise-free datasets. Existing datasets are mostly harvested from microblogging platforms such as Twitter and Reddit, relying on user annotation via distant supervision through hashtags, such as *#sarcasm*, *#sarcastic*, *#not*, etc. This method is popular since 1) only the author of a post can determine whether it is sarcastic or not, and 2) it allows large-scale dataset creation. However, this method introduces large amounts of noise due to lack of context, user errors, and common misuse on social media platforms. The sarcasm detection datasets created through manual annotation tend to be of higher quality but are typically much smaller. These problems are further compounded for non-English datasets, both manually labeled and automatically collected. Further, as sarcasm strongly relies on its context, using classical machine translation (MT) from English often produces inadequate results. This makes sarcasm detection in less-resourced languages, like Slovenian, an even bigger challenge. Therefore, developing reliable sarcasm detection models is of crucial importance for robust sentiment analysis in these languages.

We develop a methodology for sarcasm detection in less-resourced languages and test it on the Slovenian language. We address the problem of missing datasets by comparing state-of-the-art machine translation with large generative models. We explore the viability of such datasets and how the number of parameters affects a model's ability to detect sarcasm. We construct various ensembles of large pretrained language models and evaluate their performance.

The rest of this work is organized as follows. In Section 2, we discuss the proposed approach for detecting sarcasm in a less-resourced language such as Slovenian. We present the creation of a dataset, details of the training methodology and deployed ensemble techniques. We lay out our experimental results and their interpretations in Sections 2.3 and 4. In Section 5, we provide conclusions and directions for future work.

2 Sarcasm Detection Dataset

Existing attempts at automatic sarcasm detection have resulted in the creation of datasets in a small number of languages with differing sizes and quality. It is unclear if models trained on these datasets would generalize well to unseen languages [1]. Since no dataset exists for Slovenian, we leverage recent advances in machine translation and large language models (LLMs) to create a dataset for supervised sarcasm detection. We thus apply a translate-train approach when fine-tuning our models.

The prevalence of research done on sarcasm in English means that English datasets are usually larger and of higher quality than their counterparts in other languages. Further, as the translation from (and to) English is usually of better quality, we consider only English datasets.

Preliminary tests showed poor quality and poor translation ability of `Sarcasm on Reddit`[1] dataset, and `News Headlines Dataset For Sarcasm Detection`[2]. Hence, we chose the recent `iSarcasmEval`[3] dataset from the `SemEval-2022` shared task. We

[1] www.kaggle.com/datasets/danofer/sarcasm
[2] www.kaggle.com/datasets/rmisra/news-headlines-dataset-for-sarcasm-detection
[3] github.com/iabufarha/iSarcasmEval

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.4212

believe that relatively low performance scores obtained in this shared task could be improved with the use of larger LLMs.

2.1 iSarcasmEval Dataset

`iSarcasmEval` is a dataset of both English and Arabic sarcastic and non-sarcastic short-form tweets obtained from Twitter. We use only the English part, which is pre-split into a train and test set. Both sets are unbalanced, the former having 867 sarcastic and 2601 non-sarcastic examples, while the latter has 200 sarcastic and 1200 non-sarcastic examples. The authors of the shared task claim that both distant supervision and manual annotation of datasets produce noisy labels in terms of both false positives and false negatives [1]. Thus, they ask Twitter users to directly provide one sarcastic and three non-sarcastic tweets they have posted in the past. These responses are then filtered to ensure their quality. The produced dataset is not entirely clean since it contains links, emojis, and capitalized text. We chose to leave all of these potential features in the text, as they commonly occur in online conversations and could be indicative of sarcasm.

Let us mention, that an ensemble approach with 15 transformer models and transfer from three external sarcasm datasets proved to be the most accurate modeling technique for English [9] achieving an F_1-score of 0.605.

2.2 Translating iSarcasmEval

Our preliminary testing using smaller BERT-like classifiers showed that the models learned the distribution of the data and defaulted to the majority classifier (1200/1400 = 0.857 test accuracy). To try to dissuade this, we merged the train and test sets, kept all the sarcastic instances, and randomly sampled an equal number of non-sarcastic examples. This left us with a balanced dataset of 2134 samples.

To enable task specific instructions that would preserve sarcasm, we skipped classical machine translation tools, and tried two alternative translation approaches:

- using a medium-sized T5 model trained specifically for neural machine translation,
- leveraging a significantly larger model via OpenAI's API.

The T5 model uses both the encoder and decoder stacks of the Transformer architecture and is trained within a text-to-text framework. We chose Google's 32-layer T5 model called `MADLAD400-10B-MT`[4], which has 10.7 billion parameters and is pretrained on the `MADLAD-400` [4] dataset with 250 billion tokens covering 450 languages. Fine-tuning for machine translation was done on a combination of parallel data sources in 157 languages, including Slovenian.

As a generative model, we chose OpenAI's decoder-based `GPT-4o-2024-05-13`[5]. Its true size is not known to the public, but it's speculated that it is significantly smaller than `GPT-4`, since it is much faster and more efficient. OpenAI claims that it has the best performance across non-English languages of any of their models, thus making it suitable for our task.

When *prompting* generative decoder-based models, it is necessary to craft clear and specific instructions to achieve the best results. We used few-shot learning [2], and randomly sampled three training instances, manually translated them, and included them in the following prompt where the double forward slash was used as a delimiter between the query and the expected response.

You will be provided with a sarcastic/non-sarcastic sentence in English, and your task is to translate it into the Slovenian language. It should keep the original meaning. Examples:

- *love getting assignments at 6:25pm on a Friday!! // obožujem, ko mi v petek ob 18:25 pošljejo naloge!!*
- *I still can't believe England won the World Cup // Še vedno ne morem verjeti, da je Anglija zmagala na svetovnem prvenstvu*
- *taking spanish at ut was not my best decision 😐 // Učenje španščine na UT ni bila moja najboljša odločitev 😐,*

We manually assessed the outputs of both transformers in order to determine the best translations for fine-tuning detection models.

2.3 Translation Results

During translation, the T5 model sometimes had trouble with examples that had multiple newline characters in a row. It occasionally dropped parts of texts it didn't understand (mostly slang and various types of informal text styles). This shows that a 10B parameter model is not large enough to robustly translate all features of a language such as English into a less-resourced language such as Slovenian.

On the other hand, the GPT model performed surprisingly well in most instances and it seemed to have a more nuanced understanding of phrases used in online speech. It consistently translated entire texts, keeping the original structure and meaning. Consequently, we used GPT's translations when training sarcasm detection models. The translations can be seen in our repository[6].

3 Model Training

We tested the performance of a wide range of LLMs of different sizes. Their overview is contained in Table 1.

Table 1: Summary of used sarcasm detection models.

Model	Parameters
`SloBERTa`	110M
`BERT-BASE-MULTILINGUAL-CASED`	179M
`XLM-RoBERTa-BASE`	279M
`XLM-RoBERTa-LARGE`	561M
`META-Llama-3.1-8B-INSTRUCT`	8.03B
`META-Llama-3.1-70B-INSTRUCT`	70.6B
`META-Llama-3.1-405B-INSTRUCT`	406B
`GPT-3.5-TURBO-0125`	?
`GPT-4o-2024-05-13`	?

3.1 Encoder Models Under 1B Parameters

The four smallest models are encoder-based models that embed input text and use a classification head to assign it a class. They required additional fine-tuning to perform sarcasm detection. For these models, we conducted hyperparameter optimization.

We chose the `SloBERTa` [7, 8] model in order to check whether using a monolingual Slovenian model impacts sarcasm detection performance. We also wanted to compare BERT and RoBERTa-like models, so we used their multilingual variants and fine-tuned them on Slovenian data.

The models were trained for a maximum of five epochs with the use of early stopping, where the training was halted if the validation loss didn't improve after two epochs.

[4] huggingface.co/google/madlad400-10b-mt
[5] platform.openai.com/docs/models/gpt-4o
[6] github.com/GalaxyGHz/Diploma

3.2 Llama 3.1 Models

Since the teams that competed in the 2022 shared task on sarcasm mostly used BERT and RoBERTa models, we extend the testing to include significantly larger models. We chose Meta's open-source Llama family of models, more specifically, their newest Llama 3.1 variants. These come in three different sizes, which was perfect for studying the effects of parameter counts on sarcasm detection. We decided to use the "`instruct`" version of all three models since these were fine-tuned to be better at following instructions.

When prompting LLama and GPT generative models, the following few-shot classification prompt was given, with two positive and two negative examples randomly sampled from our dataset.

You will be provided with text in the Slovenian language, and your task is to classify whether it is sarcastic or not. Use ONLY token 0 (not sarcastic) or 1 (sarcastic) as in the examples:

- *Spanje? Kaj je to... Še nikoli nisem slišal za to? 1*
- *Lepo je biti primerjan z zidom* 😂 *1*
- *To sploh nima smisla. Nehaj kopati. 0*
- *Dne 12. 10. 21 ob 10:30 je bil nivo reke 0,37 m. 0.*

We used full versions of the 8B and 70B parameter models, while the 405B parameter model was loaded in 16-bit precision mode. To minimize the use of resources and costs, we employed LoRA parameter-efficient fine-tuning. We provided the models with training and validation sets and trained them for a maximum of 10 epochs. No hyperparameter optimization was conducted in this case due to time constraints. We used the validation loss to choose the best model, and we used the same early stopping technique as with the smaller models.

3.3 GPT 3 and 4 Models

We also tested two models offered on the OpenAI platform, `GPT-4o-2024-05-13` and `GPT-3.5-TURBO-0125`. We first used them in few-shot mode and classified all our examples without any fine-tuning. When fine-tuning, the platform's tier system limited us to only the smaller `GPT-3.5-TURBO-0125` model. We fine-tuned the model for a maximum of three epochs. In the end, we used the model with the lowest validation loss to classify the examples in the test set.

3.4 Sarcasm Detection Ensembles

When constructing ensemble models, we tried two techniques: stacking and voting. In both cases, we used the predicted probability of the sarcastic class from each model as input features.

3.4.1 Stacking With Regularized Logistic Regression. Our first ensemble used stacking approach, and logistic regression with Ridge regularization as the meta-level classifier. This choice was motivated primarily by the need for feature selection, as we wanted to identify the most important model predictions and determine which models would be assigned a lower weight. The best models were then used for voting.

3.4.2 Standard and Mixed Voting. The second ensembling method was voting. We tried cut-off-based mixed voting inspired by [9]. Formally, we used hard voting when the absolute difference between the number of sarcastic and non-sarcastic predictions was greater than n, and we used soft voting otherwise. We optimized the value of n based on the ensembles performance on our validation set.

When n is set to zero, this approach is equivalent to hard voting, and in the case of n being equal to the predictor count, it is equivalent to soft voting. We report both results. Additionally, we compare the results of voting using all trained models with the results obtained by using only the models with large weights in our regularized logistic regression ensemble.

4 Sarcasm Detection Results

Table 2 summarizes all our results. It is roughly sorted by model size, smaller models being on top and larger ones being on bottom. The `(NFT)` tag indicates that a model was not fine-tuned, while the `(LoRA)` tag means that a model was trained with LoRA. Results are rounded to three decimal places.

Table 2: Summary of performance results for all tested models. The best scores are in bold.

Model	Accuracy	F_1-score
`SloBERTa`	0.621	0.632
`BERT-BASE-MULTILINGUAL-CASED`	0.499	0.666
`XLM-RoBERTa-BASE`	0.578	0.579
`XLM-RoBERTa-LARGE`	0.550	0.597
`Llama-3.1-8B-INSTRUCT (NFT)`	0.560	0.676
`Llama-3.1-8B-INSTRUCT (LoRA)`	0.569	0.682
`Llama-3.1-70B-INSTRUCT (NFT)`	0.660	0.724
`Llama-3.1-70B-INSTRUCT (4-bit-LoRA)`	0.637	0.717
`Llama-3.1-405B-INSTRUCT (16-bit-NFT)`	0.686	0.751
`GPT-3.5-TURBO-0125 (NFT)`	0.564	0.679
`GPT-3.5-TURBO-0125`	0.749	0.760
`GPT-4o-2024-05-13 (NFT)`	0.686	0.746
`L2-LOGISTIC-REGRESSION`	**0.759**	**0.765**
`L2-LOGISTIC-REGRESSION-NON-COMMERCIAL`	0.707	0.749
`HARD-VOTING-ALL`	0.670	0.738
`SOFT-VOTING-ALL`	0.658	0.732
`HARD-VOTING-BEST-5`	0.686	0.749
`SOFT-VOTING-BEST-5`	0.686	0.749

Individual Model Performance

Out of all of the used models, only `BERT-BASE-MULTILINGUAL-CASED` failed to learn any pattern in our data and defaulted to the dummy classifier.

`GPT-3.5-TURBO-0125` sometimes predicts the correct token but then continues to generate additional text, such as 11 and 1\n1. This happens with a small quantity of examples in our testing set. We decided to truncate these responses and only kept the first token as the answer.

The Llama models sometimes refused to generate tokens zero or one. We decided to drop these examples altogether. We report the test accuracy and trained the ensemble models without them.

Smaller encoder models performed poorly when compared to some of the larger models. Only the `SloBERTa` model manages to achieve an accuracy above 0.6. Despite being the smallest of the four small models we tested, `SloBERTa` performed the best. This suggests that the three larger multilingual encoder models may lack sufficient understanding of Slovenian. It also highlights that model size alone does not necessarily correlate with better performance when it comes to sarcasm detection.

The Llama models fared better, achieving accuracies of up to 0.686 with the 405B model being comparable to GPT-4o in performance. They still fell short of the fine-tuned `GPT-3.5-TURBO-0125` model, which managed to successfully classify about three-quarters of our examples with a F_1-score of 0.76.

Some models had significantly higher F_1-scores and lower accuracies. We show the confusion matrix of one of the models

Table 3: Confusion Matrix for non-fine-tuned `Llama-3.1-405B-INSTRUCT` model.

Predicted \ Actual	Positive	Negative
Positive	202	123
Negative	11	91

that exhibited the largest difference in Table 3. These models seem to have a tendency to incorrectly classify non-sarcastic text as sarcastic, leading to a high rate of false positives.

Our testing also showed that loading the `Llama-3.1-70B--INSTRUCT` model in 4-bit mode and fine-tuning it with LoRA does not produce satisfactory results, and it is thus better to conduct full fine-tuning with the smaller Llama model or to use one of OpenAI's models via their fine-tuning API.

`GPT-3.5-TURBO-0125` performed the best among individual models, so if costs associated with the use of OpenAI's API are acceptable, we recommend its use for sarcasm detection in Slovenian. This shows that very large models can effectively identify sarcasm. We believe that with better parameter tuning, Llama 8B could be one of the best (and most economical) options for sarcasm detection in Slovenian, provided that the user has sufficient hardware resources.

Ensemble Model Performance

We observed that the regularized logistic regression mostly relied on the best-performing models. Its focus on the best model (`GPT-3.5-TURBO-0125`), however, suggests that there is significant overlap between the various model predictions.

We decided to discard `BERT-BASE-MULTILINGUAL-CASED` when constructing our voting ensembles since its dummy classification didn't contribute to overall model performance. Both of these two voting classifiers had an odd number of predictors, so there was no need for a tie-breaker mechanism.

Voting proved to be ineffective in our setups, even scoring lower than some of its base models. hard voting generally outperformed soft voting. We also note that there was no benefit in using mixed voting, at least for the sets of predictors that we obtained as hard voting always had a higher accuracy. This was true for both the classifiers that used all and only five of the base models.

Regularized logistic regression managed to improve on the scores of individual models, raising accuracy by one percent, thus achieving the best performance out of all of the tested approaches. This shows that there is still performance to be gained from ensembles; however, it is still necessary to use commercial models for top performance.

5 Conclusion

In this work, we presented the task of sarcasm detection in the less-resourced Slovenian language. Our code and results are freely available[7].

We tackled the missing dataset problem by using two LLMs to perform neural translation of an English dataset into Slovenian. The translations generated by `GPT-4o-2024-05-13` outpaced those generated by a large T5 model specifically trained for neural machine translation in terms of quality.

We used this data to train a plethora of Transformer-based models in various settings. We found that fine-tuning `GPT-3.5--TURBO-0125` via OpenAI's API results in the highest individual Slovenian sarcasm detection power, but we also note that a possible alternative could be local fine-tuning of the `Llama-3.1-8B--INSTRUCT` model. Our testing shows that using aggressive quantization combined with LoRA results in significant performance degradation.

We also constructed ensemble models based on voting and stacking methods. Observations showed that voting didn't result in any performance improvements. On the other hand, stacking with the use of a regularized logistic regression managed to improve on the performance of its base models.

Additional work needs to be done in dataset construction. Sarcastic examples could be extended with context or labels of the types of sarcasm they represent. This might help guide models towards better understanding of sarcasm. Future work could also explore incorporating heterogeneous models into ensembles or the creation of Mixture of Experts (MoE) ensembles, whose baseline models would focus on different aspects of sarcasm.

[7] github.com/GalaxyGHz/Diploma

Acknowledgements

This research was supported by the Slovenian Research and Innovation Agency (ARIS) core research programme P6-0411 and projects J7-3159, CRP V5-2297, L2-50070, and PoVeJMo.

References

[1] Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 802–814. DOI: 10.18653/v1/2022.semeval-1.111.

[2] Tom Brown et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*. Vol. 33, 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[3] Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic sarcasm detection: a survey. *ACM Comput. Surv.*, 50, 5, Article 73, 22 pages. DOI: 10.1145/3124420.

[4] Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. [n. d.] Madlad-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* Article 2940, 13 pages.

[5] Bleau Moores and Vijay Mago. 2022. A survey on automated sarcasm detection on twitter. *arXiv preprint*. DOI: 10.48550/arXiv.2202.02516.

[6] Smaranda Muresan, Roberto Gonzalez-Ibanez, Debanjan Ghosh, and Nina Wacholder. 2016. Identification of nonliteral language in social media: a case study on sarcasm. *Journal of the Association for Information Science and Technology*, 67, 11, 2725–2737. DOI: 10.1002/asi.23624.

[7] Matej Ulčar and Marko Robnik Šikonja. 2021. Sloberta: slovene monolingual large pretrained masked language model. In *Proceedings of Data Mining and Data Warehousing, SiKDD*, 17–20. http://library.ijs.si/Stacks/Proceedings/InformationSociety/2021/IS2021_Volume_C.pdf.

[8] Matej Ulčar and Marko Robnik Šikonja. 2021. *Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0. CLARIN.SI data & tools*. Nasl. z nasl. zaslona. Fakulteta za računalništvo in informatiko. http://hdl.handle.net/11356/1397.

[9] Mengfei Yuan, Zhou Mengyuan, Lianxin Jiang, Yang Mo, and Xiaofeng Shi. 2022. Stce at SemEval-2022 task 6: sarcasm detection in English tweets. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 820–826. DOI: 10.18653/v1/2022.semeval-1.113.

Speech-to-Service: Using LLMs to Facilitate Recording of Services in Healthcare

Maj Smerkol
maj.smerkol@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Rok Susič
rs36117@student.uni-lj.si
University of Ljubljana, Faculty of Mathematics and Physics
Ljubljana, Slovenia

Mariša Ratajec
mr97744@student.uni-lj.si
University of Ljubljana, Faculty of Electrical Engineering
Ljubljana, Slovenia

Helena Halbwachs
h.halbwachs@senecura.si
SeneCura Kliniken- und Heimebetriedsgesellschaft m.b.H.
Vienna, Austria

Anton Gradišek
anton.gradisek@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

Digital tracking of services is one of the main administrative burdens of the healthcare staff. Here, we present a proof-of-concept study of a so-called speech-to-service (S2S) system that is aimed at facilitating recording of activities, extracting information from the conversation between a healthcare provider and recipient. The system comprises of a speech recorder, a diarization component, an LLM to interpret the conversation, and a recommendation system integrated in a smart tablet that records completed activities and suggests possible other activities that may have still be required. We tested the system on 350 conversations and obtained 95% accuracy, 97% precision and 94% recall.

Keywords

healthcare, LLM, speech recognition, recommendation system

1 Introduction

Healthcare workers, including nurses, technicians, and care personnel form the backbone of the health system as they care for patients and tend to their needs. However, with the standardization and systematization of the healthcare professions and services often becomes a large bureaucratic burden, as healthcare workers have to record all the activities and services they provide to the patients. This process is of course needed as it provides traceability and ensures that all the required activities were taken care of, but the problem is that the interfaces designed for activity logging are often not user-friendly and require the users to choose the activities from a extensive lists of drop-down menus. In total, this amounts to substantial time required only for tedious administrative tasks, time that would be more beneficially spent otherwise.

With the aim to alleviate the administrative burden of activity logging, we explored the possibilities of novel technologies to assist the healthcare staff in their logging tasks. We developed and tested a proof-of-concept system that records the conversation between the healthcare worker and a patient, identifies the activities, and allows the healthcare worker to batch-confirm them on a dedicated smart tablet. Batch-confirmation saves a lot of time by significantly lowering the number of clicks required in the UI. The system is built using open-source or publicly accessible components, particularly a speech-to-text system that transcribes the recorded conversation, and a large language model (LLM) that leverages its natural language processing capabilities. The recommender system shows possible required tasks, serving as a reminder and to suggest tasks that are expected soon, which may lower the number of visits per patient. These recommended tasks are then suggested to the healthcare worker, who can review and confirm them using the LLM-assisted interface. LLMs, such as ChatGPT and Llama, have seen a surge in popularity in a wide variety of topics since their popularization in particular with the unveiling of ChatGPT3 in the autumn of 2022.

Several LLM based systems have been proposed recently, including administrative task automation [6], decision making process [10], improving existing automatic speech recognition (ASR) systems [1], and providing patients with needed information [9]. A recent study [11] concludes that utilising ASR to ease some administrative tasks leads to faster, more efficient work and even increase workers' moods.

2 System Architecture

This paper describes two early prototype systems, both aiming to alleviate the workload of healthcare workers by easing the task of documenting care actions performed. These are the ASR system that logs care actions based on captured dialogue between the healthcare worker and the patient, and a recommender system that predicts the required services at a specific time. This recommender system relies on the historical data, appropriate for long-term patient care facilities.

Both systems are limited in scope and only target the most common healthcare services in the dataset for detection or prediction respectively, which can still greatly easy the workload for medical workers, since the top 10 most common tasks out of around 200 care action types represent around 80% of all services performed.

The recommender system allows the care workers to anticipate tasks in advance and server as a reminder. This aims to lower the number of patient visits, which also alleviates the workload.

2.1 Speech-to-Service ASR

The ASR system consists of a speech diarization model, capable of segmenting the recorded speech based on who is currently speaking, a speech transcription model that transcribes the audio

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.4550

to text, and a LLM fine-tuned to extract specific information from the text. Figure 1 shows the architecture of the prototype system.

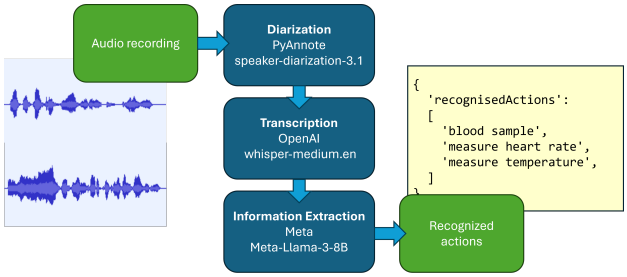


Figure 1: Overview of the ASR system.

We employ speaker-diarization pretrained model [1][4] for diarization, pyannote/speaker-diarization-3.1 pretrained model [7] for transcription, and fine-tuned Llama3 model[2][2] for information extraction via generating JSON formatted output.

2.2 Recommender System

The recommender system prototype is based on machine-learning prediction of events that are expected to occur in a certain time window for a specific patient with addition of tasks that commonly follow predicted tasks. Due to the sensitive nature of the data, we base our predictions only on the time window, patient ID and care type. Thus we consider multi-output binary classifiers that do not require large amounts of data for training. Additional tasks are added to the list based on a Markov chain model that commonly follow, e.g. the task 'clean table' follows the task 'lunch'.

The feature vector includes the time of day, day of week, week of month and month of year as numbers, allowing for capture of periodic events with different periods. Due to lack of patient data, we opted for personalized models, trained for each patient separately. We believe that results can be further improved by adding more patient-related attributes. The model training used five month period of data collected, with cross-validation, and the accuracy was evaluated on the data collected during the sixth month. Due to patients' medical states changing over time, some data drift is expected, which is reflected in our results.

3 Dataset

To fine-tune the information extraction model based on Llama3, we have created a dataset of conversations in text form and appropriate outputs for each of them, as the task on hand is very specific and we did not find any existing appropriate dataset. We automated the process and manually removed any bad examples. A real dataset, ideally recorded in the target environment, is needed for for final implementation - LLM generated datasets used for training LLMs are only appropriate in preliminary studies.

To generate the dataset, we prepared a BERT [3] LLM via prompting [5]. A training sample was generating by first randomly selecting 2 of the 10 target actions, and programmatically generating the target output JSON. The BERT model was then tasked with generating a conversation, in which these two tasks are mentioned as done during the conversation. We generated several hundred conversations that way, and manually checked for mistakes in the model output. Many conversations were removed due to selected actions not being mentioned or other reasons. Finally, the resulting dataset contains 350 conversations and JSON formatted lists of tasks.

For the prediction of services required during a visit, we have acquired a log of all services performed in one long-term patient care facility over a period of 6 months, with the next version expanding to data from six facilities. The tasks in dataset include measurements (body temperature, heart rate, blood pressure, ...), medical tasks (monitoring medicine intake, performing examinations, turning the patient in bed) and care tasks (breakfast, lunch, cleaning). There are over 200 different tasks mentioned. The dataset includes limited patient information—patient ID, care type, and a detailed chronological history of services received. Care types (CareType I, CareType II, CareType III/A, CareType III/B, CareType IIII) represent an estimate of how much assistance a person requires. Legal restrictions on accessing sensitive health data prevented us from obtaining more detailed patient records, so we developed prediction models based on these limited data points, balancing accuracy with regulatory constraints.

The data preprocessing involved determining each patient's presence in the facility by identifying the timestamps of their first and last recorded service. Patients with a stay of less than four months were excluded from the analysis to ensure sufficient data for reliable predictions.

4 Methods

This section describes the methodology used to develop the ASR system and the recommender system.

4.1 Clustering

The primary goal of the clustering process was to group patients with similar patterns in terms of the type and frequency of services they received, allowing us to predict relevant services more effectively for each cluster (since it was not clear, even among experts, whether care type and actual care provided were correlated).

The clusters, as shown in Figure 2, demonstrate that patients within the same care type tend to receive similar services. Some deviations, where multiple classifications appear within a cluster, are likely due to temporary conditions we could not fully exclude (for instance, an individual categorized under "Care Type II" may temporarily receive services typical of "Care Type III/A" (e.g. due to a broken arm), while their care type classification remains unchanged). Despite this, the care types differentiate well enough across clusters, leading us to use "CareType" as one of the key attribute for further service predictions.

In the clustering process, we excluded CareType IIII because this group is characterized by highly diverse healthcare needs due to specific diseases, and experts advised us to omit it for this part of the analysis.

4.2 Recommender System

To recommend the required services, we constructed the training dataset using a detailed log of care actions performed over a 6-month period. For each patient, the data was divided into consecutive 4-hour time windows. In each window, we examined whether specific care actions were performed, marking them as "positive" if they occurred within that time frame. This granular approach allowed us to capture the temporal dynamics of

[1] pyannote/speaker-diarization-3.1
[2] meta/meta-llama-3-8b
[3] google-bert/bert-base-multilingual-cased

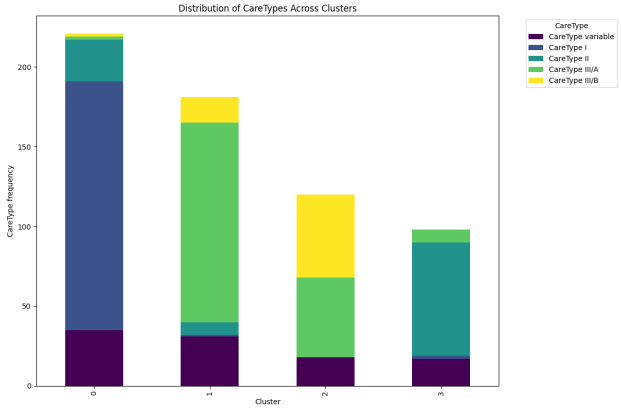


Figure 2: Clustering of patients closely aligns with pre-existing care type assignments, ranging from minimal personal assistance (CareType I) to moderate assistance (CareType II), and full or intensive personal assistance (CareTypes III/A and III/B) for those with more severe care needs.

service delivery, ensuring that for each time window, we had a clear record of the services provided. As a result, we generated over 1000 labeled examples per patient, with each example representing a specific time window and its associated care actions. This enabled the model to learn patterns in service requirements throughout the day and week.

To identify the best predictive model, we evaluated various classification algorithms, including Random Forest, Decision Tree, K-Neighbors, Support Vector Classifier (SVC), Gradient Boosting, and Naive Bayes. Each model was trained using a multi-output classification approach, with features including the frequency of the top services provided and the relevant time attributes. To ensure robust model evaluation, we implemented 5-fold cross-validation and subsequently tested the models on the sixth month's data to assess their predictive performance.

4.3 Speech Recognition and Information Extraction

Due to limited availability of training data, only the information-extraction model based on Llama3 was fine tuned using few-shot LoRA (low-rank adaptor) supervised training. The diarization and transcription models are used unchanged.

The diarization model used is speaker-diarization [4]. Initial experiments with few-shot LoRA fine tuning [3] did not improve the performance, hinting at the need for a larger training dataset. The model's performance is satisfactory at the task of segmentation, but less so at the task of identifying which segments belong to which speaker, especially for longer conversations. For a two-speaker situation, the model seems to assume the speakers take turns speaking, causing mistakes when a single speaker pauses before continuing to speak.

The transcription model used is whisper [8]. The model transcribes each segment separately. As mentioned above, the speakers are not robustly recognised, and we cannot reliably assign a speaker to each line of text. Still, labeling each line of text even with an ambiguous label improves the downstream task of information extraction. The transcribed lines of text are concatenated, and at the start of each utterance a label marking it as such is added. Thus, the transcribed text resembles a play with unknown characters speaking.

The information extraction model is Llama3 [2], and fine-tuned utilising a LoRA few-shot fine tuning. Our approach was to fine-tune the model for the task of extracting information about specific care actions and generate the output in a JSON format, providing structured data directly. A small training dataset was prepared as described in the section 3.

5 Results and Discussion

5.1 LLM Based Infromation Retrieval Model

The Llama3 based information extraction model is evaluated using a 5-fold cross validation, achieving **95% accuracy**, **97% precision**, and **94% recall**. For evaluation the model's JSON-formatted strings were deserialized to objects and tested against known correct objects to be able to interpret the results as multi-label binary classification.

The LLM infromation extraction model sometimes generates invalid JSON after fine-tuning, most commonly due to duplicated keys or getting stuck in a loop, generating same elements until maximum output size is generated. The generated strings are therefore post-processed to fix these mistakes via simple string manipulation, however this indicates that experiments with different output formats or avoiding generating the answers should be performed.

The whole ASR pipeline including diarization and transcription has not yet been evaluated and falls within the scope of future work.

5.2 Recommender System

Tables 1 and 3 present the classification results. Table 1 reports the average performance across all patients, including standard deviations for the different models, while Table 3 shows classification accuracy by care type, with averages and standard deviations across all patients within each care type, based on the model with the best results, which in this case is K-Neighbors (KNN).

Results are reported in two ways, tables 1 and 3 show accuracy considering all target attributes, only considering a prediction correct when all targets are predicted correctly. The table 2 show average of accuracies for each target attribute.

Table 1: Cross-validation and test accuracy (mean ± standard deviation) across all patients for various classification models.

Model	CV Accuracy	Test Accuracy
RandomForest	0.71 ± 0.14	0.66 ± 0.16
DecisionTree	0.65 ± 0.16	0.66 ± 0.16
KNeighbors	**0.73 ± 0.13**	**0.71 ± 0.16**
SVC	0.63 ± 0.12	0.63 ± 0.14
GradientBoosting	0.68 ± 0.12	0.66 ± 0.15
NaiveBayes	0.57 ± 0.17	0.55 ± 0.20

The K-Neighbors (KNN) classifier outperformed other models, achieving an average CV accuracy of 73%, a test accuracy of 71%, and R^2 score of 0.44. This made it the most effective model for predicting service plans. Random Forest also performed reasonably well, achieving a test accuracy of 66%, though it did not surpass KNN in overall performance.

Table 2: Majority Class Percentage and Task-wise Test Accuracy (mean ± standard deviation) across all patients for various classification models.

Model	Majority Class Percentage	Task-wise Accuracy
RandomForest	0.72 ± 0.19	0.89 ± 0.10
DecisionTree	0.72 ± 0.19	0.89 ± 0.11
KNeighbors	**0.72 ± 0.19**	**0.91 ± 0.10**
SVC	0.65 ± 0.16	0.88 ± 0.10
GradientBoosting	0.65 ± 0.16	0.89 ± 0.09
NaiveBayes	0.72 ± 0.19	0.85 ± 0.15

Table 3: Classification performance of K-Neighbors (KNN) by CareType, showing cross-validation and test accuracy (mean ± standard deviation), averaged across all patients within each care type.

CareType	CV Accuracy	Test Accuracy
CareType I	0.79 ± 0.12	0.76 ± 0.16
CareType II	0.79 ± 0.11	0.78 ± 0.13
CareType III/A	0.68 ± 0.13	0.66 ± 0.15
CareType III/B	0.70 ± 0.14	0.68 ± 0.17
CareType IIII	0.68 ± 0.10	0.67 ± 0.12

Since all predictive accuracy values exceed the 70% majority class baseline, this is an excellent result. In multi-label classification, where multiple services are predicted simultaneously, it's important to not only focus on overall accuracy but also on the accuracy of each individual task. By achieving 90% accuracy on the most common tasks, the model ensures that key services are reliably predicted.

The lower test accuracy compared to cross-validation can be explained by temporal changes in patient conditions, as the test set only included the last month of data. As patient care needs shift over time, predicting long-term patterns is more challenging than shorter-term cross-validation, where care remains more stable.

The test accuracy also reflected noticeable differences across care types. CareType I and CareType II showed higher accuracy rates, while more complex types, such as CareType III/A, III/B, and IIII, exhibited a drop in accuracy of around 10%. This is likely due to the more diverse and unpredictable care needs in these groups, making service prediction more challenging.

This approach, particularly with the strong performance of our K-Neighbors (KNN) model, demonstrated the potential of machine learning to enhance personalized planning in healthcare. In future work, including additional patient-specific features beyond time-based data, such as health-related attributes, could further improve accuracy, particularly for the more complex care types.

6 Conclusions

This is early work and further improvements are underway. The whole ASR pipeline needs to be evaluated and we expect noticeably worse performance comparing to only the information extraction model due to larger complexity and possibility for failure at each step. The information retrieval model itself is not inefficient considering computational time and memory required, but diarization and transcription steps are. The required service prediction should also be further improved. Using current dataset an alternative approach that may improve performance is using sequence modelling or event prediction approaches. Finally, the two models could work in tandem - predicting the required actions and using that information in the ASR pipeline could be beneficial.

Based on the proof-of-concept study, we conclude the suggested approach is in principle feasible and can be beneficial to healthcare providers. However, in view of regulations, special caution has to be paid during the implementation of any sort of such system in a real-world setting. Recording and diarizing conversations between healthcare staff and the patients is likely to include highly personal data, which falls under the EU relevant legislation, specifically the GDPR (*General Data Protection Regulation*) [4] and the EU AI Act (*Artificial Intelligence Act (Regulation (EU) 2024/1689)*) [5]. Furthermore, indiscriminately recording conversations and feeding them into an LLM will likely be considered as "high risk" in view of the AI Act. This means that implementing such services will require extensive screening, documentation, clear division of ownership and access roles, and other compliance with legal requirements.

Acknowledgements

We thank the healthcare provider organization for the dataset and for insightful discussions.

References

[1] Ayo Adedeji, Sarita Joshi, and Brendan Doohan. 2024. The sound of healthcare: improving medical transcription asr accuracy with large language models. *arXiv preprint arXiv:2402.07658*.

[2] AI@Meta. 2024. Llama 3 model card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[3] Shamil Ayupov and Nadezhda Chirkova. 2022. Parameter-efficient finetuning of transformers for source code. *ArXiv*, abs/2212.05901. https://api.semanticscholar.org/CorpusID:254564456.

[4] Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*. Brno, Czech Republic, (Aug. 2021).

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805. http://arxiv.org/abs/1810.04805 arXiv: 1810.04805.

[6] Senay A. Gebreab, Khaled Salah, Raja Jayaraman, Muhammad Habib ur Rehman, and Samer Ellaham. 2024. Llm-based framework for administrative task automation in healthcare. In *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, 1–7. DOI: 10.1109/ISDFS60797.2024.10527275.

[7] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. (2022). DOI: 10.48550/ARXIV.2212.04356.

[8] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. (2022). DOI: 10.48550/ARXIV.2212.04356.

[9] Prakasam S, N. Balakrishnan, Kirthickram T R, Ajith Jerom B, and Deepak S. 2023. Design and development of ai-powered healthcare whatsapp chatbot. *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, 1–6. https://api.semanticscholar.org/CorpusID:259280109.

[10] Raja Vavekanand, Pinja Karttunen, Yue Xu, Stephanie Milani, and Huao Li. 2024. Large language models in healthcare decision support: a review.

[11] Markus Vogel, Wolfgang Kaisers, Ralf Wassmuth, and Ertan Mayatepek. 2015. Analysis of documentation speed using web-based medical speech recognition technology: randomized controlled trial. *Journal of medical Internet research*, 17, 11, e247.

[4] https://gdpr-info.eu/

[5] https://artificialintelligenceact.eu/the-act/

Performance Comparison of Axle Weight Prediction Algorithms on Time-Series Data

Žiga Kolar
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
ziga.kolar@ijs.si

David Susič
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
david.susic@ijs.si

Martin Konečnik
Cestel Cestni Inženiring d.o.o
Špruha 32
Trzin, Slovenia
martin.konecnik@cestel.si

Domen Prestor
Cestel Cestni Inženiring d.o.o
Špruha 32
Trzin, Slovenia
domen.prestor@cestel.si

Tomo Pejanovič Nosaka
Cestel Cestni Inženiring d.o.o
Špruha 32
Trzin, Slovenia
tomo.pejanovic@cestel.si

Bajko Kulauzović
Cestel Cestni Inženiring d.o.o
Špruha 32
Trzin, Slovenia
bajko@cestel.si

Jan Kalin
Zavod za gradbeništvo Slovenije
Dimičeva ulica 12
Ljubljana, Slovenia
jan.kalin@zag.si

Matjaž Skobir
Cestel Cestni Inženiring d.o.o
Špruha 32
Trzin, Slovenia
matjaz.skobir@cestel.si

Matjaž Gams
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
matjaz.gams@ijs.si

Abstract

Accurate vehicle axle weight estimation is essential for the maintenance and safety of transportation infrastructure. This study evaluates and compares the performance of various algorithms for axle weight prediction using time-series data. The algorithms assessed include traditional machine learning models (e.g., random forest) and advanced deep learning techniques (e.g., convolutional neural networks). The evaluation utilized datasets comprising time-series data from 10 sensors positioned on a single lane of a bridge, with the goal of predicting each vehicle's axle weights based on the signals from these sensors. Each algorithm's performance was measured against the OIML R-134 recommendation, where a prediction was classified as accurate if the error was within ±4 percent for two-axle vehicles and ±8 percent for vehicles with more than two axles. Tests were conducted on several bridges, with this paper presenting detailed results from the Lopata bridge. Findings indicate that deep learning models, particularly convolutional neural networks, significantly outperform traditional methods in terms of accuracy and their ability to adapt to complex patterns in time-series data. This study provides a valuable reference for researchers and practitioners aiming to enhance axle weight prediction systems, thereby contributing to more effective infrastructure management and safety monitoring.

Keywords

time-series data, axle weight, machine learning, neural network

1 Introduction

Accurate axle weight prediction plays a pivotal role in the maintenance and safety of transportation infrastructure [7]. The precise estimation of axle weights is essential for various applications, including road maintenance planning, traffic management, and the prevention of overloading, which can lead to premature road wear and increased accident risks [8]. Traditional methods for axle weight measurement often rely on static scales or weigh-in-motion (WIM) systems. While these methods provide direct measurements, they are susceptible to limitations such as high installation and maintenance costs, potential measurement inaccuracies due to environmental factors, and the need for frequent calibration.

In recent years, the advent of advanced computational techniques has opened new avenues for improving axle weight prediction. Machine learning (ML) and deep learning (DL) algorithms, in particular, offer promising alternatives by leveraging time-series data to model complex, non-linear relationships inherent in vehicular weight patterns. These methods can enhance prediction accuracy, handle large volumes of data, and adapt to varying conditions, making them suitable for real-world applications where traditional methods may fall short.

This study systematically evaluates and compares the performance of various axle weight prediction algorithms using time-series data. We focus on a diverse set of algorithms, including machine learning models like random forests (RF) [6] and advanced deep learning techniques such as convolutional neural networks (CNN) [4].

The objective of this research is to explore the potential of combining traditional WIM systems with advanced ML and DL models to enhance axle weight predictions. By comparing the performance of different methodologies, including the SIWIM traditional model, random forest (IJS RF), a hybrid approach (AVERAGE(IJS, SIWIM traditional)), and a CNN-based model, this study aims to identify the most effective strategies for accurate and reliable axle weight estimation. Additionally, it examines the impact of synthetic data generation on the performance of these models, providing a comprehensive evaluation of their practical applicability in real-world scenarios.

The study aimed to predict the axle weights of vehicles using ten input signals from sensors placed under the Lopata bridge.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.4752

Each predictive algorithm's performance was evaluated according to the OIML R-134 recommendation, which is deemed accurate if the error margin for predicting the axle weight is within ±4% for vehicles with two axles and within ±8% for vehicles with more than two axles.

The dataset comprised 1478 samples, i.e. passing of a vehicle, each containing 10 signals per vehicle. For each sample, a static weight for each axle was assigned as the target value. Static weight refers to the weight measured by a scale when the vehicle is stationary.

This paper is structured as follows: Section 2 reviews several state-of-the-art approaches. Section 3 details the preprocessing steps necessary before applying machine learning methods. In Section 4, algorithms used for predicting axle weights are presented. Section 5 presents the final results of the axle weight predictions. Finally, Section 6 summarizes the findings and proposes ideas for future research.

2 Related Work

The prediction of axle weights using time-series data has often been studied in recent years, resulting in a substantial body of related work. Below, several state-of-the-art (SOTA) approaches are described.

Zhou et al. [10] differentiated between high-speed and low-speed weigh-in-motion (WIM) systems and analyzed the characteristics of axle weight signals. They proposed a nonlinear curve-fitting algorithm, detailing its implementation. Numerical simulations and field experiments assessed the method's performance, demonstrating its effectiveness with maximum weighing errors for the front axle, rear axle, and gross weights recorded at 4.01%, 5.24%, and 3.92%, respectively, at speeds of 15 km/h or lower.

Wu et al. [8] introduced a modified encoder-decoder architecture with a signal-reconstruction layer to identify vehicle properties (velocity, wheelbase, axle weight) using the bridge's dynamic response. This unsupervised encoder-decoder method extracts higher features from the original data. A numerical bridge model based on vehicle-bridge coupling vibration theory demonstrated the method's applicability. Results indicated that the proposed approach accurately predicts traffic loads without additional sensors or vehicle weight labels, achieving better stability and reliability even with significant data pollution.

Xu et al. [9] applied wavelet transform for denoising and reconstructing the WIM signal, and used a back propagation (BP) neural network optimized by the brain storm optimization (BSO) algorithm to process the WIM signal. Comparing the predictive abilities of BP neural networks optimized by different algorithms, they found the BSO-BP WIM model to exhibit fast convergence and high accuracy, with a maximum gross weight relative error of 1.41% and a maximum axle weight relative error of 6.69%.

Kim et al. [5] developed signal analysis algorithms using artificial neural networks (ANN) for Bridge Weigh-in-Motion (B-WIM) systems. Their procedure involved extracting information on vehicle weight, speed, and axle count from time-domain strain data. ANNs were selected for their effectiveness in incorporating dynamic effects and bridge-vehicle interactions. Vehicle experiments with various load cases were conducted on two bridge types: a simply supported pre-stressed concrete girder bridge and a cable-stayed bridge. High-speed and low-speed WIM systems were used to cross-check and validate the algorithms' performance.

Bosso et al. [1] proposed a method using weigh-in-motion (WIM) data and regression trees to identify patterns in overloaded truck weights and travel. The analysis reveals that truck type is the key predictor of overloading, while time of day is crucial for axle overloading, with most incidents occurring late at night or early morning. These insights can enhance enforcement strategies and inform pavement management and design, optimizing infrastructure longevity and safety.

He et al. [2] introduced a new method that uses only the flexural strain signals from weighing sensors to identify axle spacing and weights, reducing installation costs and expanding BWIM applications. The method's accuracy is validated through numerical simulations and laboratory experiments with a scaled vehicle-bridge interaction model, showing promising results for accurate axle spacing and weight identification.

3 Data Preprocessing

Before applying various algorithms to the dataset, several preprocessing steps were necessary. Due to the differing lengths of signals from each sample, padding was performed to standardize them to the length of the longest signal. Samples with a gross weight below 5 kN were excluded from both the training and test datasets. Each signal was cropped by removing data to the left of the leftmost peak value minus 100 and to the right of the rightmost peak value plus 100. The peak values were calculated in advance.

To address the limited availability of data required for deep learning, which typically necessitates tens of thousands of samples for effective training, synthetic data generation was employed. The original dataset comprised 1,478 samples (from January 2022 to December 2023) i.e. passing of a vehicle, each containing 10 signals per vehicle. An additional 20,000 synthetic samples were generated using a specific algorithm. This algorithm operates by iterating 20,000 times, during each of which a random training sample and a random strain factor were selected. The strain factor is a random value ranging between 0.5 and 0.99. The selected signal from the training sample was then scaled by the chosen strain factor. This scaling process effectively models the feature that doubling the amplitude of the signal corresponds to doubling its weight.

A crucial aspect of data preprocessing involved the normalization of sensor signals to ensure uniformity across the dataset. Each signal was normalized to have a mean of zero and a standard deviation of one, which helps in improving the convergence of machine learning algorithms by ensuring that each feature contributes equally to the learning process.

The selection of training and test data was conducted using a rolling window approach [3]. Specifically, for each testing month, the training data comprised all available data up to but not including the testing month. For instance, if May 2023 was designated as the testing month, the training dataset consisted of data from January 2022 through April 2023. This process was systematically repeated for each testing month from March 2022 to December 2023.

4 Methodology

Four methods were identified as applicable for predicting vehicle axle weights. The first method, known as SIWIM traditional [11], calculated the number of axles, axle lengths, and axle weights by utilizing influence lines to model the signal and determine the correct output. For validation purposes, each predicted output

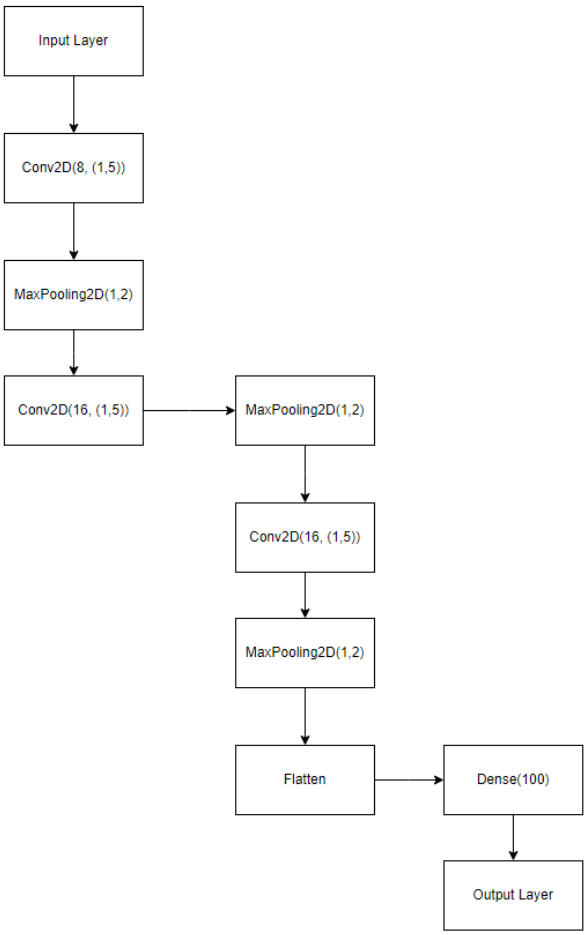


Figure 1: Architecture of CNN for predicting axle weights.

was stored in a separate file alongside the signal data, enabling direct comparison with the actual values.

The second method used the random forest [6] (named IJS RF) for predicting vehicle axle weights. The model relied on accurately identifying the positions of peaks to function correctly. Peak values were determined using the *find_peaks* method from the SciPy library, which identifies peaks based on a specified minimum height. Once the peaks were identified, the algorithm extracted values within a ±5 range of each peak. These extracted values were then used as input features for the random forest model. Additionally, the random forest model incorporated temperature, axle distances and gross weight as input features. Random forest algorithms are not inherently suited for time series data; however, they perform effectively with numerical data such as temperature, axle distance, and gross weight. Therefore, this algorithm was chosen for analyzing this type of input data.

The third method integrated the first two approaches by averaging the outputs from the SIWIM traditional and IJS RF models (named AVERAGE(IJS, SIWIM traditional)). This approach is motivated by the concept that combining multiple models can often yield more accurate results than relying on a single model alone [12].

The final method employed a convolutional neural network (CNN) to predict axle weights. The CNN utilized synthetic data, as detailed in section 3, during the training phase. This method processed all 10 signals as input to calculate the axle weights. The detailed architecture of the CNN is shown in Figure 1. 2D Convolutional layers (Conv2D) were used instead of 1D Convolutional layers due to the input data consisting of 10 sensor signals. The number of filters and kernel size are specified within the parentheses of each Conv2D layer, while the pooling size is defined in each 2D MaxPooling layer parentheses (MaxPooling2D). The last Dense layer has 100 neurons. To mitigate overfitting, a Dropout layer was added after the final Dense layer. Additionally, Batch Normalization was applied after each 2D Convolutional layer to further reduce the risk of overfitting.

Although Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) neural networks could be used for this task, a Convolutional Neural Network (CNN) was chosen instead because of its strengths in capturing spatial hierarchies and local patterns within the data. CNNs are highly effective at extracting local features and detecting patterns, while LSTM and GRU are better suited for handling temporal dependencies, which are not that relevant to this specific task.

5 Results

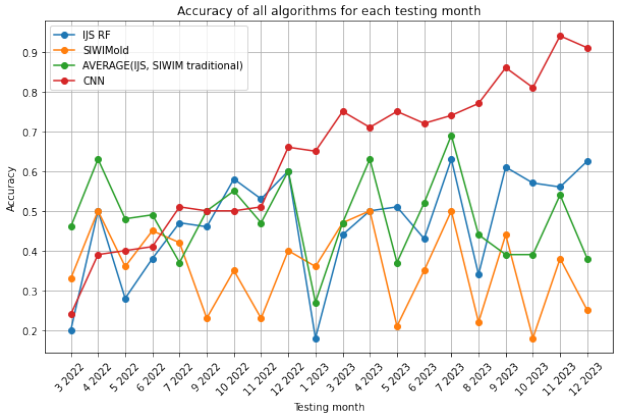


Figure 2: Accuracies of all algorithms for each testing month.

The results of each method described in Section 4 are illustrated in Figure 2. Among the methods evaluated, SIWIM traditional exhibited the poorest performance, with fluctuating trends observed throughout the entire two-year period. The CNN began to outperform the other three approaches after December 2022. Conversely, the AVERAGE(IJS, SIWIM traditional) method showed superior performance during the initial testing months from March 2022 to June 2022.

The performance of the CNN improved with an increasing amount of data, whereas the IJS RF and AVERAGE(IJS, SIWIM traditional) methods were more effective during the initial phase when less training data was available. However, the improvement in CNN's accuracy was not linear. This non-linear trend can be attributed to the random initialization of the CNN's weights before each training session, occasionally leading to suboptimal convergence.

An additional analysis was conducted to compare the performance of the models under varying environmental conditions, such as temperature fluctuations and differing traffic patterns. This analysis revealed that the CNN model maintained its accuracy more consistently across different conditions, indicating its

robustness and adaptability. Furthermore, the inclusion of synthetic data in training the CNN model contributed to its superior performance, as it allowed the model to learn from a more diverse set of examples. Future research should focus on expanding the range of synthetic data and exploring additional ensemble techniques to further enhance prediction accuracy.

Despite achieving high accuracy with the CNN model, with the highest accuracy reaching 0.94, this most accurate method still falls short of meeting the OIML R-134 recommendation by 4.4%. Furthermore, the results show that more static data could be needed for the learning phase. Having 1000 static samples which were augmented might not be sufficient to reach the OIML R-134 recommendation.

In summary, the results indicate that while traditional methods such as IJS RF and AVERAGE(IJS, SIWIM traditional) perform well with limited data, convolutional neural networks (CNNs) demonstrate superior performance as more data becomes available, despite some variability in their convergence. In addition, a sufficient number of training examples is needed to approach the desired OIML R-134 recommendation.

6 Conclusion and Discussion

In this study, a performance comparison of various axle weight prediction algorithms using time-series data collected from 10 sensors positioned on the Lopata bridge was conducted. The algorithms evaluated encompassed traditional machine learning models, such as random forests, and advanced deep learning techniques, notably convolutional neural networks.

The major findings reveal that CNNs achieved significantly better results in predicting axle weights during the latter months of the experiment. The CNNs' ability to adapt to and learn from complex patterns within the time series data was a key factor in their superior performance. Despite achieving high accuracy with the CNN model, reaching a peak accuracy of 0.94, this method still falls short of meeting the OIML R-134 recommendation by 4.4%.

Overall, there are three implications of this study. First, it demonstrates the potential of deep learning techniques to enhance the accuracy of axle weight predictions where sufficient data is available, thereby facilitating more reliable infrastructure management. Second, for smaller datasets, it is more effective to use classical machine learning systems in combination with methods like SIWIM traditional. Third, it provides a valuable benchmark for researchers and practitioners, guiding the development and implementation of more effective axle weight prediction systems.

To achieve the OIML R-134 recommendation, two options are possible:

- Just add more data - if the trend continues, adding another half a year of measurements would enable achieving the standard. Another option would be to apply measurements on a bridge with more traffic.
- Improve the methods by incorporating advanced ensemble techniques.

To introduce the ensemble approaches, one potential improvement involves modeling each sensor individually. This approach entails building a separate CNN model for each of the ten sensors, allowing for more specialized and potentially more accurate predictions from each sensor's data. By focusing on the unique characteristics and data patterns of each sensor, the models can be better tailored to capture specific nuances in the time-series data.

After developing individual models for each sensor, the next step would be to combine the predictions from these models into a single final prediction. This can be achieved using an ensemble method, such as a random forest classifier. The random forest classifier would take the ten individual predictions (one from each sensor model) as input features and produce a consolidated final axle weight prediction.

This method not only holds the potential to improve the accuracy and robustness of the axle weight predictions but also provides a scalable framework that can be adapted to different datasets and sensor configurations. Future work should explore the implementation of this approach, including the optimization of individual sensor models and the integration of their predictions through an ensemble method.

By advancing the CNN model in this manner, it is anticipated that the performance gap relative to the OIML R-134 recommendation could be further reduced, bringing the predictions closer to the required accuracy levels with a smaller amount of data and enhancing the overall efficacy of the axle weight prediction system.

Acknowledgements

This study received funding from company Cestel. The authors acknowledge the funding from the Slovenian Research and Innovation Agency (ARIS), Grant (PR-10495) and Basic core funding P2-0209. The author(s) made use of chatGPT to assist with this article. ChatGPT was commonly employed as a tool for enhancing the language of the initial draft, without altering the length of the text. ChatGPT 4 was accessed/obtained from chatgpt.com and used with modification in July 2024.

References

[1] Mariana Bosso, Kamilla L Vasconcelos, Linda Lee Ho, and Liedi LB Bernucci. 2020. Use of regression trees to predict overweight trucks from historical weigh-in-motion data. *Journal of Traffic and Transportation Engineering (English Edition)*, 7, 6, 843–859.

[2] Wei He, Tianyang Ling, Eugene J OBrien, and Lu Deng. 2019. Virtual axle method for bridge weigh-in-motion systems requiring no axle detector. *Journal of Bridge Engineering*, 24, 9, 04019086.

[3] Hamed Kalhori, Mehrisadat Makki Alamdari, Xinqun Zhu, Bijan Samali, and Samir Mustapha. 2017. Non-intrusive schemes for speed and axle identification in bridge-weigh-in-motion systems. *Measurement Science and Technology*, 28, 2, 025102.

[4] Teja Kattenborn, Jens Leitloff, Felix Schiefer, and Stefan Hinz. 2021. Review on convolutional neural networks (cnn) in vegetation remote sensing. *ISPRS journal of photogrammetry and remote sensing*, 173, 24–49.

[5] Sungkon Kim, Jungwhee Lee, Min-Seok Park, and Byung-Wan Jo. 2009. Vehicle signal analysis using artificial neural networks for a bridge weigh-in-motion system. *Sensors*, 9, 10, 7943–7956.

[6] Steven J Rigatti. 2017. Random forest. *Journal of Insurance Medicine*, 47, 1, 31–39.

[7] Mohhammad Sujon and Fei Dai. 2021. Application of weigh-in-motion technologies for pavement and bridge response monitoring: state-of-the-art review. *Automation in Construction*, 130, 103844.

[8] Yuhan Wu, Lu Deng, and Wei He. 2020. Bwimnet: a novel method for identifying moving vehicles utilizing a modified encoder-decoder architecture. *Sensors*, 20, 24, 7170.

[9] Suan Xu, Xing Chen, Yaqiong Fu, Hongwei Xu, and Kaixing Hong. 2022. Research on weigh-in-motion algorithm of vehicles based on bso-bp. *Sensors*, 22, 6, 2109.

[10] ZF Zhou, P Cai, and RX Chen. 2007. Estimating the axle weight of vehicle in motion based on nonlinear curve-fitting. *IET science, measurement & technology*, 1, 4, 185–190.

[11] A Žnidarič, J Kalin, M Kreslin, M Mavrič, et al. 2016. Recent advances in bridge wim technology. In *Proc. 7th International Conference on WIM*.

[12] Hui Zou and Yuhong Yang. 2004. Combining time series models for forecasting. *International journal of Forecasting*, 20, 1, 69–84.

Comparison of Feature- and Embedding-based Approaches for Audio and Visual Emotion Classification

Sebastijan Trojer
st5804@student.uni-lj.si
Jožef Stefan Institute
Faculty of Computer and Information Science
Ljubljana, Slovenia

Zoja Anžur
zoja.anzur@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Mitja Luštrek
mitja.lustrek@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Gašper Slapničar
gasper.slapnicar@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

This paper presents a comparative analysis of feature- and embedding-based approaches for audio-visual emotion classification. We compared the performance of traditional handcrafted features, using MediaPipe for visual features and Mel-frequency cepstral coefficients (MFCCs) for audio features, against neural network (NN)-based embeddings obtained from pretrained models suitable for emotion recognition (ER). The study employs separate uni-modal datasets for audio and visual modalities to rigorously assess the performance of each feature set on each modality. Results demonstrate that in the case of visual data NN-based embeddings significantly outperform handcrafted features in terms of accuracy and F1 score when training a traditional classifier. However, for audio data the performance is similar on all feature sets. Handcrafted features, such as facial blendshapes, computed from MediaPipe keypoints and MFCCs, remain relevant in resource-constrained settings due to their lower computational demands. This research provides insights into the trade-offs between traditional feature extraction methods and modern deep learning techniques, offering guidance for the development of future emotion classification systems.

Keywords

emotion recognition, embeddings, hand-crafted features

1 Introduction

Automated emotion recognition (ER) often focuses on two modalities – video and audio. This is akin to human emotion recognition, as we heavily rely on audio-visual characteristics, such as facial expressions and audio cues, to deduce emotional state [7]. Both audio and video are relatively simple to obtain using sensors, as such sensors are unobtrusive and easily available (e.g., webcameras) and can be used to train machine learning (ML) models for emotion recognition.

In the past decade deep-learning (DL) approaches achieved state-of-the-art (SOTA) results in many domains, including emotion recognition [16]. However, despite the superior performance of such models, many doubts have been cast on their black-box nature, lacking explainability and interpretability of the internally derived features [9]. Furthermore, while some research suggests superior performance of embeddings compared to traditional features [20], this is not universally agreed upon [8], especially when taking into account potentially much higher computational complexity of deriving embeddings with deep artificial neural networks (ANNs).

Our research question is thus, whether it is better to compute embeddings using SOTA pretrained DL models instead of using hand-crafted features, as ANN embeddings promise to increase detection accuracy at the cost of interpretability and computational complexity. In this work we compared the performance of hand-crafted features and embeddings obtained with pretrained SOTA models for the down-stream task of emotion recognition. We independently compared ER performance of audio and video modality, using established benchmark datasets for each. Handcrafted features were chosen based on literature and embeddings were computed with task-suitable pretrained models available in existing Python libraries. Both were formatted in a way that allowed us to then train a set of traditional ML models, listed in Section 3.3, for ER, using hand-crafted features, embeddings, or a union of both as inputs.

2 Related Work

Performance comparison of hand-crafted features and learned embeddings has been discussed in depth in computer vision domain. Schonberger et al. [15] demonstrated that hand-crafted features (e.g., SIFT) still perform on par or better than learned embeddings in image reconstruction. They warned of high variance across datasets when using learned embeddings as features. Similarly, Antipov et al. [2] reported similar performance of hand-crafted features (e.g., HOG) and learned embeddings when classifying pedestrian gender from images using small datasets. They also highlighted superior generalization performance of embeddings across (unseen) datasets. In emotion recognition from audio, Papakostas et al. [13] compared using hand-crafted MFCC-based features with embeddings from a custom convolutional neural network (CNN) trained on spectrograms. The latter slightly outperformed hand-crafted features by 1% on average in terms of F1 score, again showing similar performance. Ye et al. [21] recently showed that using a union of both hand-crafted features and learned embeddings achieves superior performance in user identification, compared to using each input individually.

There is moderate (but not universal) agreement in recent literature that performance between hand-crafted features and

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.6883

learned embeddings is similar, however, most work comparing their performance is limited to a single modality or task. We compared performance between two different modalities for the task of ER and investigated potential performance improvements of feature-level fusion (hand-crafted + embeddings).

3 Methodology

Our task consisted of two parts – hand-crafted features and embeddings computation, and ER model training for classification. Both were done on (separate) audio and visual modality and will be described per-modality in the following sections.

3.1 Datasets

As mentioned previously, the ER task is most-often audio-visual, so we decided to use an audio and a visual dataset to independently evaluate the performance of different feature sets. While many datasets exist that contain both modalities, they often have a problem of imprecise coarse emotion labelling [18], as labels are video-based, while emotions can be exhibited and changed in much shorter windows. Splitting video into frames yields a large number of (different) instances with the same label, so we wanted a dataset with individual image labels. As our focus was on comparing the performance of hand-crafted and embedding-based features, we chose two well-established benchmark datasets dedicated to audio and visual emotion classification. These datasets contain short audio clips and individual images with precise short-term and per-frame labels, circumventing the mentioned per-video labelling problem.

3.1.1 Audio Dataset. For evaluation on audio data we decided to use the crowd-sourced emotional multimodal actors dataset (CREMA-D) [4]. It contains short clips of 91 actors between the ages of 20 and 74 coming from a variety of races and ethnicities, who exhibited six different emotions (*Anger, Disgust, Fear, Happy, Neutral, Sad*). Each actor produced about 80 clips (small variation), saying specific sentences exhibiting different emotions. The distribution of labels was balanced, each class representing approx. 16% of the data. The intended emotions were verified with 2,443 crowd-sourced human raters as baseline. These raters predicted emotions based on audio only, video only, or both, achieving 40.9%, 58.2% and 63.6% recognition of intended (acted) emotion respectively.

3.1.2 Visual Dataset. For visual data we chose the extended Cohn-Kanade dataset (CK+) [11], which a staple dataset in ER evaluation from facial expressions. It contains images of 118 adults, aged between 18 and 50, again of different ethnicities. Participants were instructed to perform a series of 23 facial displays, relating to one of seven emotions (*Anger, Contempt, Disgust, Fear, Happy, Sad, Surprise*). The distribution of classes in CK+ is not balanced – *Surprise* is the majority class at 25% and *Contempt* the minority class at 6%, with others in between. This distribution also changes between subjects. CK+ images were reshaped to 48x48 pixels, put in grayscale format and cropped using frontal face Haar cascade classifier [1] as part of preprocessing. The emotion labels were validated by experts via facial activation unit rules (e.g., *Happy* = Activation unit 12 must be present = Lip corner puller active).

3.2 Feature Computation

For selection of hand-crafted features we relied on literature and previous work in ER for each modality. For embeddings on the other hand, we chose SOTA pretrained models trained for related tasks. We extracted embeddings at a model-specific point before the learning layers, and formatted them using principal component analysis (PCA) in order to reduce their dimensionality while maintaining the relevant information.

3.2.1 Audio Features. MFCCs are historically well-established in ER from audio [10], as they give a good approximation of the human auditory system's response. For each audio clip, we computed a common set of statistical aggregate features (averages, standard deviations) for MFCCs, Root Mean Square (RMS) energy (volume), Zero-Crossing Rate, Spectral Bandwidth, Spectral Contrast, and Spectral Roll-off, using the `librosa` python library.

For embeddings we decided to investigate models pretrained on similar audio tasks (e.g., emotion recognition) and use them to the point where embeddings are available, which typically means the upper part of the ANN architecture, responsible for computation of embeddings representing the features. Three pretrained models were investigated in our evaluation, all based on the wav2vec2 architecture, which is a self-supervised model for learning speech representations proposed by Facebook AI Research (FAIR) [3]. Full wav2vec2 pretraining framework comprises a latent feature encoder, a context network using the transformer architecture, a quantization module and contrastive loss (pre-training objective). For our purposes the feature encoder is important, which is a 7-layer 1D CNN reducing the dimensionality of audio inputs into a sequence of feature vectors. The initial model version was pretrained on the LibriSpeech dataset, another version was fine-tuned on IEMOCAP dataset specifically for ER, and finally a large general cross-lingual model (XLSR) was trained on millions of hours of unlabeled audio data in 53 (later extended) languages [5]. These three variants were used to extract their corresponding embeddings. Since the input data from CREMA-D is of inconsistent shape (varying by < 1 sec), we had to employ an additional adaptive average pooling layer to ensure consistently shaped outputs. We designed this pooling layer so that we lost minimal information (short segment length for pooling) and the outputs were then flattened. PCA was employed to subsequently reduce them to 10 dimensions. The number of dimensions was chosen arbitrarily and could be changed, however, we believe that 10 dimensions offer a good balance between retained information and computational (and spatial) requirements. Moreover, this number of PCA components is on the same order of magnitude as the number of hand-crafted features, making them more comparable.

3.2.2 Visual Features. For visual features, we focused on the movement of specific facial keypoints, such as the corners of the mouth and eyebrows, which form the basis of the Facial Action Coding System (FACS) – a taxonomy that categorizes human facial expressions based on muscle movements [6]. We employed the MediaPipe (MP) framework [12] to extract values representing the activation of various facial blendshapes, which correspond approximately to the regions defined in FACS. In this paper, we classify MediaPipe features as "handcrafted" because, despite being neural network-based, they quantify predefined facial areas with human-interpretable metrics. This contrasts with CNN-based embeddings, which capture patterns without direct interpretability.

For comparison, we used embeddings from two pretrained models: FaceNet [17] and EfficientNet [19] from the HSEmotion library [14]. FaceNet architecture is based on GoogleNet, which is a variant of deep CNN, and is trained using triplet loss. It

was optimized for facial recognition, verification, and clustering. EfficientNet comprises several inverted bottleneck convolutional residual blocks. It achieved SOTA results on the AffectNet ER dataset, while being relatively light-weight. Again, PCA was used to reduce the embeddings to 10 dimensions.

3.2.3 Computational and Spatial Requirements. In order to have a clear overview of the trade-off between computational and spatial requirements of each feature computation method, and their classification performance discussed in the next section, we first report the average times to compute and disk sizes of the output (per one instance) for each method in Table 1.

Table 1: Average time and disk space needed for feature computation using each method.

Modality	Feature method	Avg. Time	Avg. Space
Audio	MFCC stats	**19 ms**	**< 1 kB**
	wav2vec2 LibriSpeech	99 ms	194 kB
	wav2vec2 XLSR	274 ms	258 kB
	wav2vec2 IEMOCAP	101 ms	5 kB
Video	MediaPipe	10 ms	**< 1 kB**
	FaceNet	29 ms	2 kB
	EfficientNet	**2 ms**	5 kB

When interpreting the results in Table 1, it must also be considered that DL-based methods require additional computational time when doing PCA on top of the raw embeddings.

3.3 Emotion Classification

Data splitting is a crucial step in evaluation of ML models, as it must be done in a way to avoid overfitting and provide a robust evaluation of generalization capabilities of a model. The aim of this research was primarily not to evaluate the absolute performance of ER, but rather compare the performance when using hand-crafted vs. embedding features. Therefore it was crucial to consistently ensure that the same data splits and models were used in each experiment, for each of the compared inputs. We decided for the most robust leave-one-subject-out (LOSO) evaluation, always using default model hyperparameters. Such experimental setup minimized overfitting and also gave a good overview of generalization performance of emotion classifiers.

4 Experiments and Results

The outputs of the previous step were used as inputs (features) to train a traditional ML model for emotion classification. We evaluated several options: taking the 10 PCA components of embeddings obtained from each pretrained model as inputs, taking only hand-crafted features as inputs, and taking union of both as input. Each of these cases was evaluated for audio and visual modality separately, using the LOSO experimental setup. Several popular ML models were compared (with default hyperparameters), including k-nearest Neighbours (kNN), Random Forest (RF), Support Vector Machines (SVM) with linear kernel, and eXtreme Gradient Boosting (XGB). We monitored classification accuracy and macro F1 score as metrics of the model performance. All results were compared with baseline majority classifier and are reported as averages across all iterations of LOSO, where majority was always taken from the train data (all except left-out).

4.1 Audio Emotion Classification

As mentioned in Section 3 we investigated the following options as feature inputs:

(1) Hand-crafted statistical features relating to MFCCs
(2) 10-component PCA of wav2vec2 embeddings from a model trained on LibriSpeech
(3) 10-component PCA of wav2vec2 embeddings from a model trained on IEMOCAP
(4) 10-component PCA of wav2vec2 embeddings from a cross-lingual XLSR model
(5) Union of hand-crafted and best-performing embeddings (from above)

These were compared in experiments as described in Section 3.3, using a set of four ML models. Results of best-performing model for each set in terms of accuracy and F1 are given in Table 2. Fused data was acquired by concatenating the feature sets.

Table 2: Best performing models for each feature set and corresponding accuracy and F1 scores for audio data. Note that embeddings were represented with 10 components obtained from PCA.

Feature set	Best model	Accuracy	F1 score
N/A	Majority	0.17±0.00	0.05±0.00
MFCC stats	RF	0.46±0.08	0.43±0.09
wav2vec2 LibriSpeech	SVM	0.47±0.08	0.45±0.09
wav2vec2 XLSR	SVM	0.30±0.05	0.27±0.05
wav2vec2 IEMOCAP	SVM	0.47±0.08	0.42±0.09
MFCC + **best** wav2vec2	SVM	**0.52±0.09**	**0.50±0.10**

4.2 Image Emotion Classification

To stay consistent with the audio experiments we performed the same LOSO experiments described in Section 3.3. We compared model performances using the following features as inputs:

(1) MediaPipe blendshapes
(2) 10-component PCA of FaceNet embeddings
(3) 10-component PCA of EfficientNet embeddings
(4) Union of MP and FaceNet embeddings
(5) Union of MP and EfficientNet embeddings

Accuracy and F1 scores for the best performing models for each set of features are again reported in Table 3

Table 3: Best-performing models for each feature set and corresponding accuracy and F1 scores for visual data. Note that embeddings were represented with 10 components obtained from PCA.

Feature set	Best model	Accuracy	F1 score
N/A	Majority	0.25±0.00	0.40±0.00
MediaPipe	RF	0.62±0.28	0.51±0.29
FaceNet	SVM	0.45±0.30	0.36±0.30
EfficientNet	RF	**0.93±0.16**	**0.90±0.20**
Mediapipe + FaceNet	XGB	0.70±0.28	0.60±0.29
Mediapipe + EfficientNet	XGB	**0.93±0.17**	**0.90±0.21**

4.3 Discussion

From Tables 2 and 3 we can observe that for audio the best performance is achieved when using union of hand-crafted and embedding features, while for visual ER the performance of only embeddings or union is nearly identical. The improvement of feature union is thus generally small, as for visual data we get the same result as using only the best embeddings (1% difference in standard deviation), while for audio data the improvement in

both metrics is about 5% compared to individual feature sets. All results substantially outperform the baseline majority classifiers.

For audio data we can see that the best embedding set (wav2vec2 LibriSpeech) performs nearly the same as hand-crafted features (MFCC stats), which is in agreement with some literature [13]. It is surprising that LibriSpeech embeddings slightly outperform IEMOCAP ones, since the latter were trained specifically for emotion recognition, while the former were not. The subpar performance of XLSR is expected, since it is a more general cross-lingual unsupervised model, while investigated data is spoken in English. For visual data on the other hand the best embeddings (EfficientNet) substantially outperform hand-crafted facial expression features (MediaPipe) and those obtained from FaceNet. This is expected, as EfficientNet was trained specifically for emotion recognition, while FaceNet was trained for face recognition. In terms of ML models, we consistently observed best performance of SVM for ER from audio data, while for video data the best model is not as homogeneous. Importantly, performance of different models (RF, SVM and XGB) was often within 1%.

Another important observation is the relative stability of results across subjects when classifying from audio, with standard deviations around 8%. The same was not observed in the evaluation from visual data, with much higher standard deviations, indicating lower stability and greater variation between subjects.

To address our initial research question, we observed similar performance of hand-crafted features and embeddings from SOTA DL models for audio-based ER, with union of both achieving the best results. The image-based visual ER achieved much better performance with learned embeddings as inputs, while the union of features showed no improvement. However, the cost of hand-crafted features and embeddings in terms of computational power required to compute, and spatial requirements to save, is not the same. While hand-crafted features are usually computed quickly and represented with a few numbers, as reported in Table 1, the embeddings require loading a (commonly large) pretrained ANN, which performs a large number of matrix multiplications, resulting in high-dimensional embeddings (e.g., 64×512). This in turn requires additional dimensionality reduction, such as PCA employed in this work. Our results indicate that for image-based visual ER, the additional cost is worthwhile, due to large improvements in performance, while audio-based ER achieved much smaller improvement, making the use of embeddings from pretrained models less attractive.

Finally, hand-crafted features mostly offer direct interpretability (e.g., audio loudness), while embeddings are commonly black-box in nature, lacking explainability without suitable mechanisms on top. The clear meaning of hand-crafted features can be helpful when training traditional ML models, where feature importance can be compared and subsequently interpreted.

5 Conclusion

In summary we compared using hand-crafted features, embeddings of pretrained SOTA models, or union of both, as inputs for ER models using audio and visual data. We found that embedding-based approach is substantially superior with visual data, outweighing the computational cost – the latter is in fact the lowest when using EfficientNet. For audio data, the improvement was only seen in union of inputs, and was relatively low.

As future work it would be worthwhile to compare merged audio-visual features and embeddings in a single ER problem on the same dataset having both modalities. Furthermore, currently used data was simulated/acted, so interpretation of these results must take that into account. Numbers are expected to decrease on a more realistic dataset, as emotions in everyday life are quite subtle [18]. It would thus make sense to run similar experiments on more realistic data as well, although such data is more scarce.

Acknowledgements

This work was supported by bilateral Weave project, funded by the Slovenian Agency of Research and Innovation (ARIS) under grant agreement N1-0319 and by the Swiss National Science Foundation (SNSF) under grant agreement 214991.

References

[1] Shahad Salh Ali, Jamila Harbi Al' Ameri, and Thekra Abbas. 2022. Face detection using Haar cascade algorithm. In *2022 Fifth College of Science International Conference of Recent Trends in Information Technology (CSCTIT)*, 198–201. DOI: 10.1109/csctit56299.2022.10145680.

[2] Grigory Antipov, Sid-Ahmed Berrani, Natacha Ruchaud, et al. 2015. Learned vs. hand-crafted features for pedestrian gender recognition. In *Proceedings of the 23rd ACM international conference on Multimedia*, 1263–1266.

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, et al. 2020. Wav2vec 2.0: a framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33, 12449–12460.

[4] Houwei Cao, David G Cooper, Michael K Keutmann, et al. 2014. Crema-d: crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5, 4, 377–390.

[5] Alexis Conneau, Alexei Baevski, et al. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

[6] Paul Ekman and Erika L Rosenberg. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

[7] Monica Gori, Lucia Schiatti, and Maria B. Amadeo. 2021. Masking emotions: face masks impair how we read emotions. *Frontiers in Psychology*, 12, 669432.

[8] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 2022. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35, 507–520.

[9] Xuhong Li, Haoyi Xiong, Xingjian Li, et al. 2022. Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, 64, 12, 3197–3234.

[10] MS Likitha, Sri Raksha R Gupta, K Hasitha, et al. 2017. Speech based human emotion recognition using MFCC. In *2017 international conference on wireless communications, signal processing and networking (WiSPNET)*, 2257–2260.

[11] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, et al. 2010. The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 94–101.

[12] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, et al. 2019. Mediapipe: a framework for building perception pipelines. (2019). https://arxiv.org/abs/1906.08172.

[13] Michalis Papakostas, Evaggelos Spyrou, Theodoros Giannakopoulos, et al. 2017. Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition. *Computation*, 5, 2, 26.

[14] Andrey Savchenko. 2023. Facial expression recognition with adaptive frame rate based on multiple testing correction. In *Proceedings of the 40th International Conference on Machine Learning (ICML)* (Proceedings of Machine Learning Research). Vol. 202. Pmlr, (July 2023), 30119–30129. https://proceedings.mlr.press/v202/savchenko23a.html.

[15] Johannes L Schonberger, Hans Hardmeier, Torsten Sattler, et al. 2017. Comparative evaluation of hand-crafted and learned local features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1482–1491.

[16] Liam Schoneveld, Alice Othmani, and Hazem Abdelkawy. 2021. Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Recognition Letters*, 146, 1–7.

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: a unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, (June 2015). DOI: 10.1109/cvpr.2015.7298682.

[18] Gašper Slapničar, Zoja Anžur, Sebastijan Trojer, et al. 2024. Contact-free emotion recognition for monitoring of well-being: early prospects and future ideas. In *Intelligent Environments 2024: Combined Proceedings of Workshops and Demos & Videos Session*. IOS Press, 58–67.

[19] Mingxing Tan and Quoc V. Le. 2019. Efficientnet: rethinking model scaling for convolutional neural networks. *CoRR*. http://arxiv.org/abs/1905.11946.

[20] Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, et al. 2021. Welfake: word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8, 4, 881–893.

[21] Cuicui Ye, Jing Yang, and Yan Mao. 2024. Fdhfui: fusing deep representation and hand-crafted features for user identification. *IEEE Transactions on Consumer Electronics*.

Multi-modal Data Collection and Preliminary Statistical Analysis for Cognitive Load Assessment

Ana Krstevska
Department of Intelligent Systems
Jožef Stefan Institute
Ljubljana, Slovenia
ana.krstevska2001@gmail.com

Sebastjan Kramar
Department of Intelligent Systems
Jožef Stefan Institute
Ljubljana, Slovenia
sebastjan.kramar@ijs.si

Hristijan Gjoreski
Faculty of Electrical Engineering and Information Technologies
Skopje, Macedonia
hristijang@feit.ukim.edu.mk

Martin Gjoreski
Università della Svizzera italiana (USI)
Lugano, Switzerland
martin.gjoreski@usi.ch

Junoš Lukan
Department of Intelligent Systems
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
junos.lukan@ijs.si

Sebastijan Trojer
Department of Intelligent Systems
Jožef Stefan Institute
Ljubljana, Slovenia
st5804@student.uni-lj.si

Mitja Luštrek
Department of Intelligent Systems
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
mitja.lustrek@ijs.si

Gašper Slapničar
Department of Intelligent Systems
Jožef Stefan Institute
Ljubljana, Slovenia
gasper.slapnicar@ijs.si

Abstract

To mitigate distractions during complex tasks, ubiquitous computing devices should adapt to the user's cognitive load. However, accurately assessing cognitive load remains a significant challenge. This study aims to present sophisticated, multi-modal data collection, which can enable accurate estimation of cognitive load using wearable and contact-free devices. A total of 25 participants participated in six cognitive load-inducing tasks, each presented at two levels of difficulty. Simultaneously, physiological and behavioral data were collected from a multi-modal sensory setup, including: Empatica E4 wristband, Emteq OCOsense glasses, an eye tracker, a thermal camera, a depth camera and an RGB video camera. Additionally, participants provided subjective measures of cognitive load by completing standardized NASA Task Load Index (NASA TLX) and Instantaneous Self-Assessment (ISA) questionnaires following each cognitive task. Preliminary statistical analyses were conducted on participant demographics, performance metrics, and the perceived difficulty of tasks, as reported in the completed questionnaires.

Keywords

cognitive load inference, wearable sensors, contact-free unobtrusive sensors

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.6961

1 Introduction

Human attention is a critical resource that is increasingly targeted by mobile applications, online services, and other forms of digital engagement. In an era of constant connectivity, capturing and retaining user attention has become a primary objective for many technologies. However, as users engage in cognitively demanding tasks, distractions can lead to performance degradation and increased stress. Therefore, to minimize interruptions and maintain productivity, ubiquitous computing systems must become capable of recognizing and adapting to the user's cognitive load in real time.

Cognitive load, defined as the mental effort required to process information and perform tasks, triggers a series of physiological responses in the human body. These responses are largely governed by the activation of the sympathetic nervous system. When cognitive load increases, measurable changes can be observed in physiological markers, including blood pressure, brain activity, eye movements, electrodermal activity (EDA), respiration rate, heart rate variability, etc. Furthermore, changes are also reflected in facial expressions, posture, and other behavioural patterns.

This study seeks to offer a unique multi-modal dataset with a rich set of wearable and unobtrusive sensors to capture the subtle changes that occur with the gradual activation of the sympathetic nervous system. Rather than solely focusing on maximizing data accuracy through the use of numerous devices, this approach also aims to identify the minimum set of sensors required to achieve reliable cognitive load assessment. To that end, rich multi-modal data was collected from a myriad of sensors, including wearables

(OCOsense glasses and Empatica E4 wristband) and contact-free unobtrusive sensors such as an advanced eye tracker, a thermal camera, a depth camera, and an RGB video camera. To the best of our knowledge, no prior dataset exists containing such rich multi-modal data obtained with such an elaborate sensory setup.

2 Related Work

The challenge of cognitive load estimation has been extensively studied across various fields. Significant emphasis has been placed on reducing cognitive load in dynamic environments, such as aviation [1]. Recent research has increasingly focused on transitioning from direct measurements, such as electroencephalography (EEG), to indirect methods of cognitive load assessment. For instance, ocular metrics, including pupil diameter and blink rate, have been shown to accurately estimate cognitive load [2, 3, 4]. Additionally, facial temperature variations have been widely correlated with cognitive workload, providing another non-invasive means of assessment [5, 6]. Novak et al. demonstrated that biometric indicators, such as galvanic skin response and skin temperature, can signal increased cognitive load; however, these measures are insufficient to distinguish between varying levels of cognitive load [7]. Wang et al. demonstrated that visual cues—including face pose, eye gaze, eye blinking, and yawn frequency—can serve as indicators of cognitive load [8].

This research aims to address the complexities of cognitive load estimation by integrating a wide range of psychophysiological signals, offering a more comprehensive approach to this task.

3 Experimental Setup

The objective of our data collection was to capture participants' cognitive load under varying levels of difficulty imposed by cognitive load-inducing tasks. The study was conducted in a quiet, temperature-controlled room, with participants tested individually. At the beginning of each session, participants were seated in a comfortable chair in front of a 24” monitor and given instructions about the experiment and their expected role. The Empatica E4 wristband was then fitted to the participant's non-dominant hand, and the OCOsense glasses for emotion recognition were equipped in line with product instructions.

Data collection was further enriched through the use of unobtrusive sensing technologies, including a Tobii Spark eye tracker (60 frames per second), an Intel RealSense Depth Camera D455 (providing depth data at 30 fps), a Logitech BRIO stream 4k webcam at 10 fps with HDR and noise-canceling microphones and a FLIR Lepton 3 thermal camera delivering a full 160x120 pixel thermal resolution with 8 fps. We used this set of devices to continuously monitor participants throughout the recording session. The experimental setup can be observed in Figure 1.

4 Data Collection Protocol

Prior to the experiment, participants completed a brief sleep questionnaire to gather information about their sleep patterns (e.g., hours slept the night before and usual sleep duration) and rated their levels of fatigue and focus on a scale of 1 to 10.

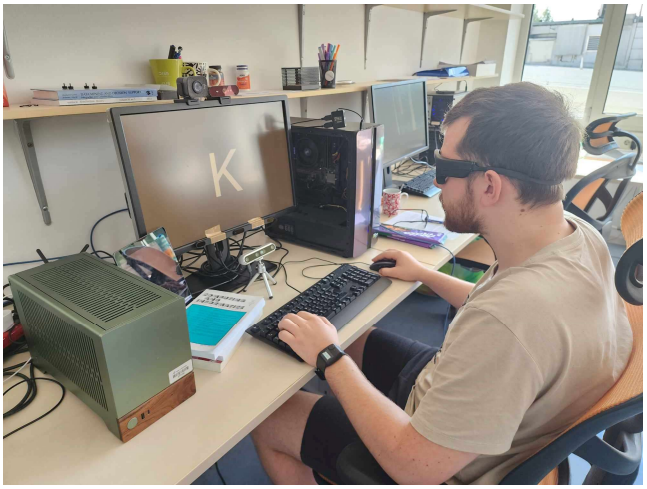


Figure 1: Experimental setup

Calibration data for the OCOsense glasses was then recorded by having participants replicate four facial expressions — smiling, frowning, brow raising, and eye squeezing — three times each. Calibration for the eye tracker followed, during which participants tracked a moving dot with their eyes. This calibration aimed to optimize participant's seating position for accurate eye-tracking.

The experiment's main phase involved participants completing cognitive load-inducing tasks that tested three aspects of cognition: attention, memory, and visual perception. For each cognitive domain, two widely recognized tasks were presented, each with two levels of difficulty (easy and difficult). This design allowed for the differentiation of cognitive load levels. Following each category of cognitive tasks, participants engaged in relaxation tasks that were not expected to induce cognitive load, such as meditation with open eyes, listening to music to relieve stress and passive viewing of aesthetically pleasing images. These tasks provided baseline data for periods of minimal cognitive load.

In summary, each recording session included six cognitive load-inducing tasks (with two levels of difficulty) and three relaxation tasks, totaling 15 tasks. After each task, participants completed the NASA Task Load Index (NASA TLX) questionnaire, a validated instrument for assessing cognitive load across six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration [9]. Each question was rated on a scale of 0 to 100. In this study, the unweighted version of the NASA TLX, known as the Raw NASA TLX, was used. Additionally, participants completed a single-item Instantaneous Self-Assessment (ISA) of workload, providing a subjective measure of the cognitive load induced by the task [10]. These questionnaires served as subjective assessments of cognitive load and as reference points for the difficulty of each task [11].

The tasks were implemented using PsychoPy, an open-source software package commonly used in neuroscience and experimental psychology research [12]. For attention-related tasks, participants completed the N-back and Stroop tests. In the N-back task, participants were presented with a sequence of letters and asked to determine whether the current letter matched the one

presented N trials earlier (with task difficulty increasing as N increased) [13]. Participants completed both a 2-back and a 3-back task. In the Stroop test, participants identified whether the word matched the color in which it was written, with the easier version involving two colors (red and blue) and the more difficult version incorporating five colors [14].

Memory-related tasks included a memory game and a question-answering task based on a previously shown image. In the memory game, participants recalled as many words as possible from a set, with the easier version comprising seven words and the more difficult version consisting of 15 words. In the question-answering task, participants focused on an image and then answered questions about it (e.g., remembering the number of particular objects in the image), with the hard version using an image with greater detail.

The visual perception tasks included a "spot the difference" task and a pursuit test. In the "spot the difference" task, participants were presented with two images and were asked to identify as many differences as possible within a one-minute time frame. The difficulty of this task varied, with the more challenging version involving an image that contained greater detail compared to the simpler, easier version. The pursuit test required participants to visually track irregularly curved, overlapping lines. As with the "spot the difference" task, the pursuit test was administered at two levels of difficulty. The more difficult version featured a more intricate image, with longer and more tangled lines, as opposed to the less complex image used in the easier version of the task.

5 Statistics

In this section, we present some descriptive demographic and task-related statistics for the participants involved in the experiment. The average age of participants was 29.28 years, with a standard deviation of 8.31. In terms of educational background, the majority of participants (44 %) had obtained a Bachelor's degree (BSc), followed by those with a Master's degree (MSc), 28 %. A smaller portion had completed only high school (16 %) or had earned a PhD (12 %). Additionally, 60 % of the participants were male.

We then looked at the descriptive statistics derived from the performance of the participants in each task. These indicate that participants performed consistently well on tasks such as the 2-back task, both easy and difficult versions of the Stroop test, the easy memory task (where participants recalled an average of 5 out of 7 words), the easy version of the "spot the difference" task (with an average detection rate of approximately 90 % of all the differences), and both versions of the pursuit test. Notably, participants performed slightly better on the difficult version of the Stroop test, likely due to their increased familiarity with the task.

However, performance was lower on tasks such as the 3-back test (which most participants perceived as highly or extremely difficult), the difficult memory task (with an average recall rate of 39 %), and both the easy and difficult question-answering tasks. The difficult version of the "spot the difference" task also showed lower performance, with participants detecting only 25 % of the differences on average. Consistent performance among subjects (with low standard deviation) was observed across all tasks except for the N-back tasks. Notably, the N-back tasks were always presented first to participants, suggesting that they may have required additional time to adjust to the testing environment and fully engage with the task.

Next, an inferential statistical analysis was performed on the relationship between task scores and various variables of the sleep pattern. To investigate the potential influence of tiredness on performance, responses from the sleep patterns questionnaire were analyzed. A non-parametric Kruskal-Wallis test was performed to determine whether there was a statistically significant difference in overall scores across different levels of tiredness (low, medium, and high). The resulting p-value (0.91) indicated no significant difference in performance between these groups. Thus, tiredness levels did not show a statistically significant impact on performance within a 95 % confidence interval.

Similarly, the effect of focus level (low vs. high) on overall performance was examined using a non-parametric Mann-Whitney test. The p-value was 0.12, indicating no statistically significant difference in performance between low and high focus groups at the 5 % significance level.

Furthermore, the relationship between hours of sleep the night before the experiment and participant performance was examined using Spearman's correlation. The p-value was 0.42, indicating no statistically significant correlation between overall performance scores and hours of sleep the night before the experiment.

The potential influence of participants' highest education level on overall performance was also investigated. To assess this, a non-parametric Kruskal-Wallis test was conducted. The results (p-value of 0.33) indicated no statistically significant difference in performance scores across different education levels among the participants.

Overall, the small sample size may have constrained the ability to detect significant effects. The limited variability in the sample's educational background and other factors likely contributed to the lack of observed differences, emphasizing the need for a larger, more diverse sample to better understand the impact of these variables on cognitive load performance.

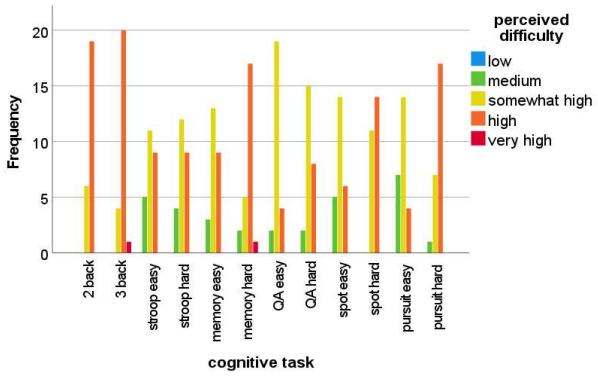


Figure 2: Reported perceived difficulty per cognitive task

As shown in Figure 2, participants consistently perceived the difficulty of the two N-back tasks and the difficult version of the "spot the difference" task as somewhat high or high. This suggests a general consensus regarding the difficulty of these tasks. In contrast, the NASA TLX-based perceived difficulty of remaining tasks, exhibited significant variability among participants.

To assess differences in performance across task difficulties and evaluate the potential for differentiating cognitive load using machine learning models, we conducted additional inferential statistical analyses. The Wilcoxon signed-rank test was used to compare participant performance on the easier and more difficult versions of each cognitive task.

Statistically significant differences in performance were found between the two difficulty levels for all tasks. For the N-back, "spot the difference", and pursuit tasks, participants performed significantly better on the easier versions, indicating that increased difficulty negatively impacted performance. Conversely, for the Stroop, memory, and question-answering tasks, participants performed better on the more difficult versions.

The statistical analysis conducted in this study provides initial evidence supporting the validity of the data collection protocol, particularly with respect to the selection of tasks and task difficulty levels. The tasks chosen for this experiment varied significantly in terms of their cognitive demands, as reflected by the substantial differences in performance between the easier and more difficult versions of each task. These results indicate that cognitive load and performance are task-specific, and the significant differences observed support the feasibility of using machine learning models to differentiate between varying levels of cognitive load.

6 Conclusion and Future Work

This study employs a novel approach to data collection for cognitive load inference by combining psychophysiological data obtained from multi-modal sensory setup, including wearable and unobtrusive contact-free sensors. The decision to utilize a diverse set of devices was motivated by the hypothesis that integrating data from multiple sources could provide a more accurate assessment of cognitive load, while also aiming to identify the minimal sensor configuration required to achieve reliable results. This is particularly relevant in dynamic and high-stakes environments, such as driving, where accurate cognitive load assessment could have life-saving implications. To the best of our knowledge, no prior research has incorporated such a comprehensive and multifaceted setup for cognitive load evaluation.

The statistical analyses conducted thus far offer promising validation for the data collection protocol. The selection of tasks and task difficulty levels proved effective in eliciting a range of cognitive load levels, as evidenced by the significant performance differences between task difficulties.

To further enhance the validity of the data collection protocol, several changes could be implemented in potential subsequent collections. Refining task difficulty levels could offer more granularity in cognitive load differentiation, ensuring a clearer distinction between varying levels of cognitive load. Furthermore, increasing the diversity of participants in terms of age, educational background, and other demographic factors is desirable to enhance the generalizability of the findings.

In future work, the collected data will be processed and utilized to train machine learning models aimed at estimating cognitive load. Ground truth for the machine learning models can be derived from various sources, including perceived task difficulty reported through the standardized questionnaires, the designed difficulty level of the tasks or the participants' performance on the tasks. These machine learning models will leverage sophisticated ML techniques to effectively integrate and analyze multi-modal data, aiming to enhance the accuracy of cognitive load predictions. We also plan to further expand the dataset with another phase of data collection, offering a rich dataset both in terms of modalities, as well as in terms of participants. The collected dataset will serve as a stepping stone towards robust multi-modal cognitive load assessment, allowing for creation and benchmarking of ML models and will be made available to general public after the collection is finalized.

Acknowledgements

This work was supported by the Jožef Stefan Institute and Università della Svizzera italiana (funded by SNSF through the project XAI-PAC (PZ00P2_216405)).

References

[1] Jonathan Mead, Mark Middendorf, Christina Gruenwald, Chelsey Credlebaugh, and Scott Galster. 2017. Investigating Facial Electromyography as an Indicator of Cognitive Workload. In *19th International Symposium on Aviation Psychology*, 377–382.

[2] Muneeb Imtiaz Ahmad, Ingo Keller, David A. Robb, and Katrin S. Lohan. 2020. A framework to estimate cognitive load using physiological data. *Personal and Ubiquitous Computing*, 27, 2027–2041.

[3] Tobias Appel, Christian Scharinger, Peter Gerjets, and Enkelejda Kasneci. 2018. Cross-subject workload classification using pupil-related measures. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 4, 1–8.

[4] Tobias Appel, Natalia Sevcenko, Franz Wortha, Katerina Tsarava, Korbinian Moeller, Manuel Ninaus, Enkelejda Kasneci, and Peter Gerjets. 2019. Predicting Cognitive Load in an Emergency Simulation Based on Behavioral and Physiological Measures. In *Proceedings of the 21st ACM International Conference on Multimodal Interaction (ICMI)*, 154–163.

[5] Fangqing Zhengren, George Chernyshov, Dingding Zheng, and Kai Kunze. 2019. Cognitive load assessment from facial temperature using smart eyewear. In *Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, 657–660.

[6] Yomna Abdelrahman, Eduardo Velloso, Tillman Dingler, Albrecht Schmidt, and Frank Vetere. 2017. Cognitive Heat: Exploring the Usage of Thermal Imaging to Unobtrusively Estimate Cognitive Load. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 33, 1–20.

[7] Klemen Novak, Kristina Stojmenova, Grega Jakus, and Jaka Sodnik. 2017. Assessment of cognitive load through biometric monitoring. In *7th International Conference on Information Society and Technology (ICIST)*.

[8] Zixuan Wang, Jinyun Yan, and Hamid Aghajan. 2012. A framework of personal assistant for computer users by analyzing video stream. In *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*, 1–3.

[9] Sandra G. Hart, and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*, 52, 139-183

[10] Andrew J. Tattersall, and Penelope S. Foord. 2007. An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics*, 39, 740-748.

[11] Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W. Woźniak. 2023. A Survey on Measuring Cognitive Workload in Human-Computer Interaction. *ACM Computing Surveys*, 55, 1–39.

[12] Jonathan Peirce, Rebecca Hirst, and Michael MacAskill. 2022. Building Experiments in PsychoPy. Sage Publications

[13] Michael J. Kane, and Andrew Conway. 2016. The invention of n-back: An extremely brief history. *The Winnower*

[14] John Ridley Stroop. 1992. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 121, 15-23

Predicting Health-Related Absenteeism with Machine Learning: A Case Study

Aleksander Piciga
ap7377@student.uni-lj.si
Faculty of Computer and Information Science,
University of Ljubljana
Ljubljana, Slovenia

Matjaž Kukar
matjaz.kukar@fri.uni-lj.si
Faculty of Computer and Information Science,
University of Ljubljana
Ljubljana, Slovenia

Abstract

Health-related absenteeism, or sick leave, is a complex issue with significant financial and operational implications for businesses. We present a machine learning approach to predict employee absenteeism in a Slovenian company. The study involved preprocessing and augmenting the dataset by incorporating domain knowledge, and evaluating various machine learning models. Gradient Boosted Regression Trees emerged as the most effective model, significantly outperforming the baseline model which merely predicted the previous year's absenteeism rate. Key attributes influencing absenteeism were identified, notably including current absenteeism, performance evaluations, and various job type and location-related features. Results highlight the potential of machine learning in proactively managing absenteeism and offer recommendations for future research, such as modeling absenteeism as a time series and incorporating additional data sources. We also show that the current data is not detailed and granular enough to further improve the results.

Keywords

absenteeism, data analysis, data augmentation, machine learning

1 Introduction

Absenteeism — temporary absence from work due to health reasons — is awidespread issue. In Slovenia, it has been on the rise since 2014 (Figure 1), with an average of 56,128 individuals absent daily in 2022, representing approximately 5.91% of the workforce [8]. This carries substantial financial burdens: direct costs like sick pay and indirect costs from overstaffing, reduced productivity and service quality [2]. The complexity of absenteeism, rooted in personal and organizational factors, makes it challenging to predict and manage effectively [10].

Recent years have witnessed a growing interest in leveraging artificial intelligence (AI) and machine learning (ML) to address the absenteeism challenge [5]. Various machine learning techniques, including neural networks, decision trees, random forests, and gradient boosting, have been employed to predict absenteeism and identify its underlying causes [3, 9]. These studies have demonstrated the potential of machine learning in providing valuable insights for proactive absenteeism management.

This paper presents a case study conducted in collaboration with a Slovenian IT company[1] aiming to improve absenteeism prediction and management. The study includes preprocessing and augmenting the company's employee data by incorporating domain knowledge, and evaluating various machine learning models. The findings highlight key attributes influencing absenteeism and offer recommendations for future research and interventions.

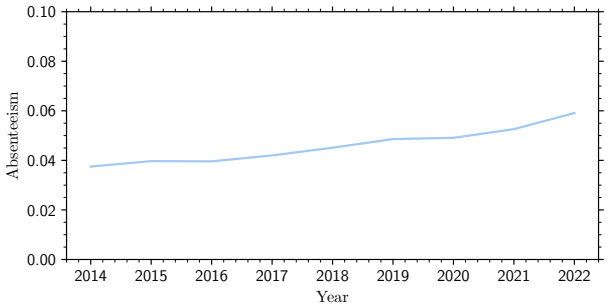


Figure 1: The increase in absenteeism rate in Slovenia between 2014 and 2022 [8]. We can observe a steady increase throughout the years.

The significance of our work extends beyond Company X, offering a blueprint for organizations tackling absenteeism. By showcasing machine learning's efficacy in predicting absenteeism and revealing its drivers, we contribute to the broader field and pave the way for data-driven interventions promoting a healthier, more productive workforce. This aligns with the growing trend of using AI and ML to address complex organizational challenges. Insights from such analyses can aid in strategic workforce planning, optimize resource allocation, and ultimately contribute to a more sustainable and resilient organization.

In section 2 we detail the data and preprocessing, section 3 outlines the methodology, section 4 presents the results, and section 5 discusses the findings and concludes the study.

2 Materials

The data used in our work spanned six years, from 2017 to 2022, and initially comprised 13,798 instances (aggregated employee records) with up to 49 attributes each. They include demographic details, work-related factors, performance evaluations and the current year's absenteeism rate for each employee, but no particulars about sick leave and other personal data.

The initial dataset required substantial preprocessing to prepare it for analysis and machine learning [6]. The data cleaning process involved addressing inconsistencies in attribute values, such as removing extraneous spaces and converting text to lowercase for uniformity. A significant challenge in the dataset was the presence of missing values, denoted by '/'. Their meaning and handling were discussed with a company representative to determine their origins and ensure appropriate treatment. In some

[1]The company asked to remain anonymous, so it is referred to as Company X.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.7260

cases, missing values were imputed based on the average values of similar instances. For example, missing values in 'Kilometers to work' were attributed to errors in data entry and were imputed using the average value for employees living in the same location and working at the same place. On the other hand, missing values in performance evaluations were due to employee's absence on evaluation days.

The target variable — health-related absenteeism rate in the following year — is a continuous variable ranging from 0 to 1. It signifies the proportion of workdays an employee is absent due to health reasons compared to the total number of workdays. The distribution of this target variable is heavily skewed to the right, with most values clustered near zero, indicating that the majority of employees have low absenteeism rates. However, there exist some outliers with extremely high absenteeism rates (Figure 2).

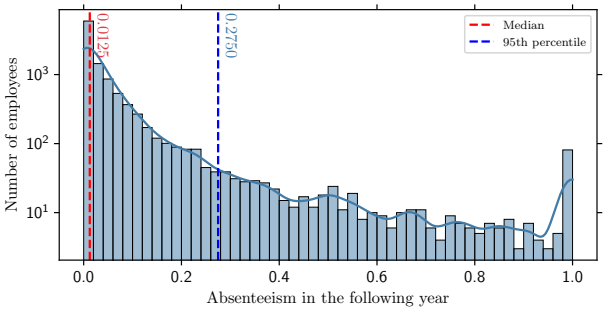


Figure 2: Log-distribution of the target variable. Most workers have very little absence, causing a right-tailed distribution with an "outlier" spike on the right.

The skewed distribution of the target variable has implications for the statistical analysis and machine learning modeling. Therefore, non-parametric statistical tests, such as the Spearman's rank correlation and Kruskal-Wallis test, were employed in EDA and data preprocessing. Additionally, the presence of outliers necessitates careful consideration during model building and evaluation.

The final dataset, comprising 10,347 instances and 42 attributes, serves as the foundation for the subsequent machine learning, where various models are trained to predict absenteeism rates.

3 Methods

The research methodology encompassed a multi-faceted approach, integrating exploratory data analysis, feature engineering, and the application of diverse machine learning models. The ultimate goal was to establish a robust predictive framework for health-related absenteeism, while also ensuring model interpretability to observe actionable insights.

3.1 Exploratory Data Analysis (EDA)

The initial phase involved a thorough EDA to understand the underlying data distribution, identify potential outliers, and uncover preliminary relationships between attributes and the target variable (absenteeism in the following year). Given the skewed nature of the target variable, visualizations like histograms and box plots were complemented by non-parametric statistical tests. The Spearman's rank correlation coefficient was employed to assess monotonic relationships between continuous attributes and the target variable, while the Kruskal-Wallis test was utilized to discern statistically significant differences across groups defined by categorical attributes.

3.2 Data augmentation/Feature Engineering

The original dataset underwent a series of transformations to enhance its suitability for machine learning. This included data cleaning, handling missing values, and the creation of new attributes based on domain knowledge and insights from the EDA. New attributes were engineered based on domain knowledge and statistical analysis. These included indicators for elevated absenteeism, receipt of bonuses or awards, high and low performance evaluations, and absenteeism rates within the employee's team and job type. External factors, such as average absenteeism rates in the employee's residential and work locations, were also incorporated. The feature engineering process was iterative, involving close collaboration with domain experts to ensure the derived attributes were meaningful and captured relevant aspects of employee behavior and organizational dynamics.

3.3 Machine Learning Models

Several well-known machine learning models were employed for absenteeism prediction, including Decision Trees, Linear Regression with L1 regularization, K-Nearest Neighbors (KNN), Support Vector Regression (SVR), Gradient Boosted Regression Trees (GBRT), and Random Forest. Hyperparameter optimization was conducted by using Optuna toolkit [1] to optimize model performance.

3.4 Model Evaluation and Selection

Model evaluation was performed using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2). The models were trained on past years' data and tested on the subsequent year, with the training set size increasing each year. The MAE provided an intuitive measure of the average prediction error, while the RMSE penalized larger errors more severely. The R2 quantified the proportion of variance in the target variable explained by the model. The models were also compared against a baseline model that simply predicted the previous year's absenteeism, to gauge the added value of the machine learning approach. A baseline model predicting the previous year's absenteeism rate was used for comparison.

3.5 Model Interpretation

SHAP (SHapley Additive exPlanations) values [4, 7] were calculated to interpret model predictions and assess attribute importance. SHAP values provide insights into the contribution of each attribute to the model's output, aiding in understanding the factors driving absenteeism. SHAP values provide a unified framework for interpreting any machine learning model, quantifying the contribution of each feature to the model's prediction for a given instance. By analyzing the SHAP values, it was possible to identify the most influential attributes and their directional impact on absenteeism.

3.6 Data Splitting

To ensure robust model evaluation and mitigate the risk of overfitting, the dataset was split into training and testing sets in a prequential manner (year after year). The models were trained on the training set and their performance was assessed on the unseen testing set for the subsequent year. This comprehensive methodological framework enabled a systematic exploration of

the factors influencing health-related absenteeism and the development of a predictive model to proactively manage this critical issue.

4 Results

The primary objective of our work was to develop machine learning models capable of predicting health-related absenteeism in the subsequent year. The models were evaluated using three key metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). The baseline model, which simply predicted the previous year's absenteeism, served as a benchmark for comparison (Table 1).

Table 1: Model performance averaged year-over-year.

Model	RMSE	MAE	R^2
Random Forest	0.107	0.052	0.344
GBRT	0.108	0.051	0.333
Linear Regression	0.108	0.051	0.331
Regression Decision Tree	0.112	0.051	0.281
KNN	0.121	0.057	0.173
SVR	0.117	0.075	0.215
Baseline Model	0.121	0.051	0.156

As we can see, all machine learning models outperform the baseline model in terms of RMSE and R^2. This indicates their superior ability to explain the variance in the target variable (absenteeism in the following year). While the MAE remains relatively consistent across models, the improvement in RMSE and R^2suggests that the models are particularly effective in handling larger deviations in absenteeism predictions.

To establish the statistical significance of the model improvements, we conducted a paired T-test comparing the predictions of each model against the baseline model. All the selected models demonstrated statistically significant improvements ($p < 0.05$) in RMSE and R^2; this ensures that their superior performance is statistically substantiated and not due to chance.

4.1 Performance Trends and Impact of Additional Data per Employee

To gain deeper insights into model behavior, we examined their performance trends over the years. Figure 3 illustrates the evolution of MAE for each model.

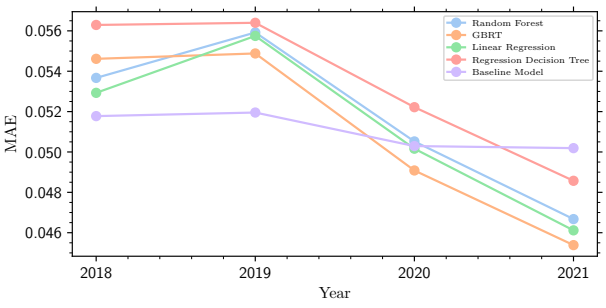


Figure 3: MAE trend over time with additional training data from past years.

Among the evaluated models, GBRT exhibited the best performance, achieving an MAE of 0.045, RMSE of 0.10, and $R^2$of 0.40 on the latest year's data. These results were statistically significantly better than the baseline model, demonstrating the effectiveness of GBRT in capturing the complex patterns underlying absenteeism.

Figure 4 reveals a general trend of MAE improvement for most models in later years, surpassing the baseline model in the final year. This suggests that the models benefit from the increasing amount of training data available in later years. RMSE and R^2charts (not shown) exhibit almost identical properties. It is clear that ML models profit tremendously from increasing amounts of data, as can be expected.

Given the observed performance gains in later years with larger training sets, we explored the impact of incorporating data from previous years. Figure 4 showcases the change in MAE for the final year when models were trained on data from the past year and the past three years, respectively.

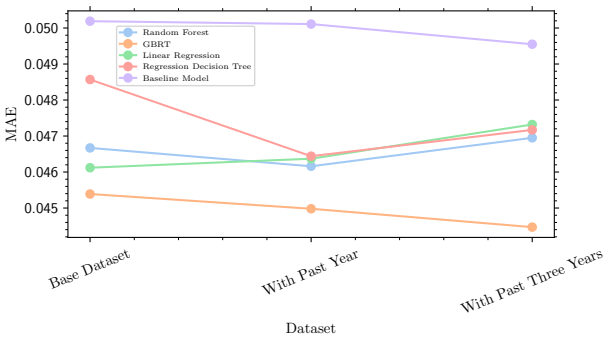


Figure 4: Impact of additional attributes from past years on MAE.

The GBRT model exhibited notable improvement with the inclusion of additional data, achieving an MAE of 0.044, RMSE of 0.093, and R^2 of 0.36. This underscores the value of historical data in enhancing the predictive capabilities of machine learning models for absenteeism and suggests that including even more historical data per employee would be beneficial.

4.2 Interpretability and Additional Insights

Analysis of SHAP values yielded the following key attributes influencing absenteeism:

- Current absenteeism rate
- Performance evaluations
- With respect to the employee's job type and location:
  - Absenteeism rate
  - Proportion of employees with elevated absenteeism
  - Proportion of employees without bonuses

Our findings suggest that absenteeism is influenced by a combination of individual factors (current absenteeism, performance evaluations) and organizational factors (job type, location, bonuses).

Additionally, a rather simple EDA visualisation of functional grouping of employees was quite surprising (Figure 5). Its interpretation can be quite speculative, possibly related to increased job satisfaction or engagement in certain groups. Another, somewhat surprising finding from EDA is that the COVID-19 pandemic did not significantly influence absenteeism rates in 2020, but it may have in 2021 (Figure 6).

Finally, t-SNE visualization of the full dataset shows that employees cannot easily be separated in clusters with similar absenteeism (Figure 7). We can identify some distinct subgroups

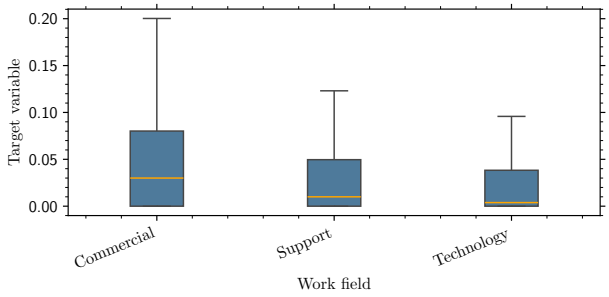


Figure 5: Target variable according to functional partitioning within the company.

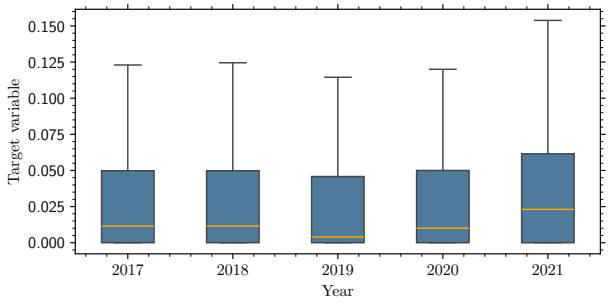


Figure 6: Target variable by year. Note the sharp increase in 2021, possibly attributable to the COVID-19 pandemic.

(like the cluster of red dots on the left), however most data points are quite intermingled. This suggests that with our current set of attributes, we shouldn't anticipate a significant improvement in predictive performance.

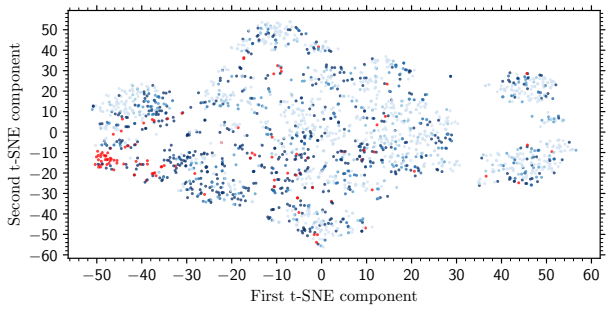


Figure 7: Data visualized in 2D space with t-SNE projection. Red dots represent examples with absenteeism in the next year above 0.25. Blue shades depict examples with absenteeism between 0 (light blue) and 0.25 (dark blue).

5 Discussion and Conclusion

Our work successfully demonstrates the application of machine learning to predict health-related absenteeism. The GBRT model's superior performance highlights its ability to capture complex data relationships, outperforming simpler models and the baseline. Also, identifying key attributes influencing absenteeism, such as current absenteeism, denied bonuses, work type and location, and performance evaluations, provides valuable insights.

The findings align with existing literature highlighting the multifactorial nature of absenteeism. The strong influence of current absenteeism on future absenteeism emphasizes its predictive power, suggesting that past behavior can be a significant indicator of future trends. The negative correlation between performance evaluations and absenteeism suggests that employees with higher evaluations tend to be less absent, potentially due to increased job satisfaction or engagement. The impact of denied bonuses on absenteeism points to the potential role of financial incentives and recognition in influencing employee attendance.

The limitations of our work include the relatively short time span and the potential influence of unmeasured external factors. Future research could address these limitations by: modeling absenteeism as a time series to capture its dynamic nature, incorporating additional data sources such as employee surveys, participation in wellness programs, and (within legal limits) health and personal circumstances data analyzing absenteeism at a finer granularity (e.g., monthly or daily), exploring the inclusion of employee health records and workplace environmental factors in predictive models, and conducting longitudinal studies to track absenteeism patterns over extended periods.

While quantitative improvements of ML model predictions are not overwhelming, the gained insights can enable targeted interventions to reduce absenteeism and promote a healthier workforce. By leveraging ML and data-driven insights, organizations can proactively manage absenteeism, thus improving productivity, financial stability, and employee well-being.

Acknowledgements

The authors sincerely thank to Company X for providing the data, domain expertise and several fruitful discussions.. The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P2-209).

References

[1] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. 2019. Optuna: a next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 2623–2631. ISBN: 9781450362016. DOI: 10.1145/3292500.3330701.

[2] M. Bregant, E. Boštjančič, J. Buzeti, M. Ceglar Ključevšek, A. Hiršl, M. Klun, T. Kozjek, N. Tomaževič, and J. Stare. 2012. *Izboljševanje delovnega okolja z inovativnimi rešitami*. Združenje delodajalcev Slovenije.

[3] B. Hu. 2021. The application of machine learning in predicting absenteeism at work. In *2021 2nd International Conference on Computing and Data Science (CDS)*, 270–276. DOI: 10.1109/CDS52072.2021.00054.

[4] Y. Meng, N. Yang, Z. Qian, and G. Zhang. 2021. What makes an online review more helpful: an interpretation framework using xgboost and shap values. *Journal of Theoretical and Applied Electronic Commerce Research*, 16, 3, 466–490. DOI: 10.3390/jtaer16030029.

[5] I. H. Montano, G. Marques, S. G. Alonso, M. López Coronado, and I. de la Torre Díez. 2020. Predicting absenteeism and temporary disability using machine learning: a systematic review and analysis. *Journal of Medical Systems*, 44, 9, (Aug. 2020), 162. DOI: 10.1007/s10916-020-01626-2.

[6] A. Piciga. 2024. *Napovedovanje zdravstvenega absentizma s strojnim učenjem*. Bachelor's Thesis. Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=slv&id=160413.

[7] E. Štrumbelj and I. Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41, 3, (Dec. 2014), 647–665. DOI: 10.1007/s10115-013-0679-x.

[8] M. Zaletel, D. Vardič, and M. Hladnik. 2024. Zdravstveni statistični letopis Slovenije 2022. (2024). Retrieved June 5, 2024 from https://nijz.si/publikacije/zdravstveni-statisticni-letopis-2022/.

[9] W. Zaman, S. Zaidi, A. I. Abdullah, and B. Touhid. 2019. Predicting absenteeism at work using tree-based learners. In *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing*. Association for Computing Machinery, 7–11. DOI: 10.1145/3310986.3310994.

[10] S. Zupanc. 2011. *Absentizem, kolegialnost in obremenjenost posameznikov*. Bachelor's Thesis. Univerza v Ljubljani. http://www.cek.ef.uni-lj.si/UPES/zupanc1175.pdf.

Puzzle Generation for Ultimate Tic-Tac-Toe

Maj Zirkelbach
mz5153@student.uni-lj.si
University of Ljubljana, Faculty of Computer and Information Science
Ljubljana, Slovenia

Aleksander Sadikov
aleksander.sadikov@fri.uni-lj.si
University of Ljubljana, Faculty of Computer and Information Science
Ljubljana, Slovenia

Abstract

Ultimate Tic-Tac-Toe is an interesting and popular variant of Tic-Tac-Toe that lacks available resources for improving gameplay skills. In this paper, we present a semi-automatic system for generating puzzles as a part of a larger tutorial application aimed at teaching Ultimate Tic-Tac-Toe. The puzzles are designed to enhance players' tactical and strategic understanding by presenting game scenarios where they must identify correct continuations. To ensure the quality of generated puzzles we tested the application with a group of volunteers. The results have shown that the number of solved puzzles positively impacted users' ability to reach higher strength levels but had less of an effect on lower levels.

Keywords

Ultimate Tic-Tac-Toe, puzzle generation, minimax algorithm, tutor application

1 Introduction

For centuries, people have enjoyed playing board games like chess and Go. Over time, these games have led to the development of extensive theory and the accumulation of knowledge, helping players navigate their complexity. Today, advanced artificial intelligence (AI) programs such as AlphaZero [14] surpass even the strongest human players, offering new insights into strategies. However, many lesser-known games have yet to be thoroughly explored, despite their popularity. One such game is Ultimate Tic-Tac-Toe, an advanced version of the classic Tic-Tac-Toe. This game is played on a 3x3 grid of local Tic-Tac-Toe boards, creating a global board (Figure 1a). The goal is to win three local boards in a row, while players must make their moves within specific local boards determined by their opponent's previous move. For example, if a player moves in the top-left corner of a local board, the next player must play on the top-left local board. If the designated board is full or decided, the player can choose any other available board. Despite its apparent simplicity, the game has enough spatial complexity that it cannot currently be solved using brute-force methods.

While there are several online implementations of the game, most focus on building strong AI agents; however, There is a noticeable lack of resources aimed at teaching and helping players understand the deeper strategies of the game, which could make the learning curve more manageable for new and aspiring players. Thus, we have created an application that addresses the lack of learning tools available for Ultimate Tic-Tac-Toe. This article places particular emphasis on the puzzle generation aspect of our application, which is designed to enhance players' tactical and strategic thinking.

In Section 2 we present the related work, and in Section 3 we detail the technical aspects of the developed application. In Section 4 we present the implemented agents and their approximate strength. In Section 5 we provide a description of different types of puzzles and the methodology for their construction. In Section 6 we present the evaluation and discuss the results in Section 7. Finally, in Section 8 we present the conclusions and give possible extensions and enhancements for future work.

2 Related Work

There are many implementations of the Ultimate Tic-Tac-Toe available online, mostly appearing as mobile games aimed primarily at entertainment and lacking advanced playing agents [12] [9] [10], as well as web and desktop applications developed to create the strongest possible programs [15] [7] [13]. An example of the latter is an agent based on the ideas of the AlphaZero program [14], currently considered one of the strongest players of this game [13]. During the development of this agent, significant strategies were discovered, which were also useful in developing our application. Some researchers have attempted to solve the game theoretically, but the spatial complexity proved too great to allow for a complete solution [5].

It is important to differentiate between the various versions of Ultimate Tic-Tac-Toe. One variant allows the game to continue playing on already-won local boards, which drastically changes the game's dynamics. In this variant, researchers have demonstrated an optimal strategy for the starting player, who can win in 43 moves [1]. Further research has focused on enabling a more balanced game by introducing random opening moves, which reduces the predictability of forced wins [4]. Despite these interesting findings, research on these variants is not so relevant for us, as it does not contribute to the understanding of the main game.

While there is a lack of educational material specific to our game, much can be learned from related fields, such as chess, which has been extensively researched. The paper by Gobet and Jansen [8] describes a scientific approach to learning chess, which includes methods to improve memory, perception, and problem-solving skills in players. In this context, it focuses on the acquisition and organization of knowledge, including both explicit and implicit learning of tactics and strategies. This approach facilitates a deep understanding of games and the development of more effective learning methods.

Chess also offers highly sophisticated practical solutions from which we can learn a great deal. Platforms such as `chess.com` [2] and `lichess.org` [11] offer extensive resources and tools for learning chess, especially in the areas of tactics and openings. These platforms allow players to learn through interactive lessons, solving puzzles, and studying various openings, which contribute to a deeper understanding of the game and improve playing skills.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.7299

This approach has proven extremely effective in helping players master complex strategic and tactical concepts in chess.

On the mentioned platforms, the methods for learning tactics are designed to allow players to solve problems based on concrete game situations, which improves pattern recognition and decision-making abilities in real games. Similarly, learning openings involves demonstrating optimal opening moves and their continuations, helping players develop effective strategies at the beginning of the game.

We have applied similar methods in our Ultimate Tic-TacToe application. For example, adapting approaches for learning tactics can help users improve their recognition and solving of complex situations in the game while learning openings helps to understand key opening moves and their impact on the further course of the game. By incorporating these methods into our application, we ensured more effective learning processes and improved the overall gaming experience.

3 Application Details

In addition to puzzle-solving, the app offers a comprehensive learning experience through various other features. It includes AI opponents of different difficulty levels, game analysis, and exploration of effective opening strategies, allowing players to refine their understanding in all phases of the game. The user interface ensures smooth navigation between these modes, making the app a versatile tool for both playing and learning Ultimate Tic-Tac-Toe. By integrating these elements, the app serves as a resource for players at all levels, helping them to deepen their understanding and improve their skills.

To reach a broader audience, the application was developed for both Android and Windows, the dominant operating systems in the market [15]. It uses Flutter components to deliver a responsive and user-friendly interface. Local data storage is utilized for user settings, progress, and puzzle data, ensuring efficient performance and data management.

We employed modern technologies and mobile development practices, including state management patterns, to create an easily expandable app for future updates and enhancements. The entire project is hosted on GitHub, though it is not open-source.

Test versions of the app for Android and Windows are available on Google Drive: https://drive.google.com/drive/folders/1SnO_mN_ZVa2wXd0OGI07kLiYKQTDHuEe?usp=drive_link, while the Android production version is accessible on Google Play Store: https://play.google.com/store/apps/details?id=com.uttt_tutor.

4 AI Agents and Rating System

Playing against intelligent agents allows users to refine their skills by competing against various virtual opponents. The application includes nine different agents, each varying in difficulty and gameplay strategies. These agents are designed using Minimax and Monte Carlo Tree Search [3] algorithms, which provide different levels of complexity and depth in move analysis. The agents and their approximate strengths are shown in Table 1.

To better understand the quality of the agents and evaluate user progress, we need to establish a system for measuring their strength. Since Ultimate Tic-Tac-Toe is not widely popular, there is no established system for rating player abilities. Therefore, we decided to use the chess rating system as an approximation for our agents.

The chess rating system is used to measure the playing strength of chess players. The most commonly used system is the Elo rating [6], which predicts the likelihood of one player winning against another based on their ratings:

$$E_A = \frac{1}{1 + 10^{\frac{R_B - R_A}{400}}},$$

where E_A represents the expected score for player A, R_A is the rating of player A, and R_B is the rating of player B.

Table 1: Table of approximate agent strengths. Each agent played 100 games (50 as X and 50 as O) against the agent one level lower. The results column shows the number of points each agent earned with each symbol, as well as the total score. A win awarded 1 point, while a draw awarded 0.5 points. The last line shows the results of the strongest freely available agent against level 9. It had the same amount of time to think, and they played 30 games.

Agent	Result			Estimated Rating
	X	O	Combined	
Confused Chimp - 1	-	-	-	1
Goofy Goblin - 2	49	49	98	620
Casual Carl - 3	41.5	35.5	77	835
Average Joe - 4	37	25	62	926
Hustling Hugo - 5	39.5	34.5	74	1114
Witty Walter - 6	43	30	73	1293
Thinking Tiffany - 7	35	24	59	1361
Brainy Bob - 8	42,.5	26.5	69	1506
Bossman - 9	36.5	22.5	59	1574
UTTT AI	14.5	12.5	27	1948

5 Puzzle Description and Methodology

In this section, we describe different types of puzzles and the methodology employed to generate them for our game.

5.1 Puzzles

The puzzles in the application are divided into tactical and strategic, with each type of puzzle covering different aspects of the game and helping players improve specific skills.

Tactical puzzles are useful for understanding tactical ideas and are particularly applicable in the endgame and middlegame phases. They focus on specific situations that require precise and thoughtful moves, helping players develop the ability to think quickly and effectively. In total, we generated 1,263 tactical puzzles, distributed across five levels. The number of puzzles for each level is shown in Table 2.

Unlike tactical puzzles, strategic puzzles aim to understand the position and long-term plans. They are instrumental in the opening and middlegame, where it is crucial to recognize strategic ideas and develop plans that provide an advantage as the game continues. There are 50 strategic puzzles available, currently arranged in one level, with the possibility of expansion in the future.

5.2 Tactical Puzzle Generation

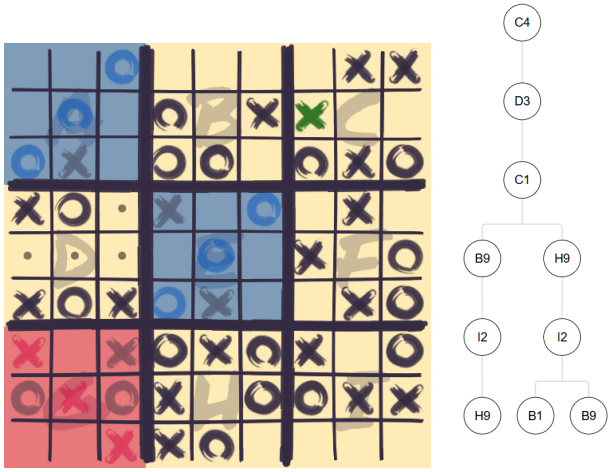
To generate tactical puzzles, we developed a specialized minimax agent that builds a tree of all possible moves leading to victory from the solver's perspective. A key step in this process is the

Table 2: Number of tactical puzzles on each level.

Level	Puzzle Depth	Quantity
1	1	273
2	3	493
3	5	231
4	7	176
5	9	90

selection of tree branches to retain only relevant and correct solutions. It is essential to preserve all of the winner's possibilities while limiting the loser's responses to those that make finding a solution as difficult as possible. Therefore, we select the continuation that allows the longest possible game for the loser while leading to the fewest continuations for the winner.

From the tree, we extract all the correct solutions for the given position. For a high-quality puzzle, it must not have too many solutions. The criterion we set is that the number of solutions must be less than the depth of the puzzle. We also decided to discard all puzzles that have multiple correct continuations for the first move. This way, we avoid trivial puzzles that would be too simple. An example of a level 3 tactical puzzle with its generated solution tree is shown in Figure 1.



(a) Level 3 tactical puzzle. **(b) Solution tree.**

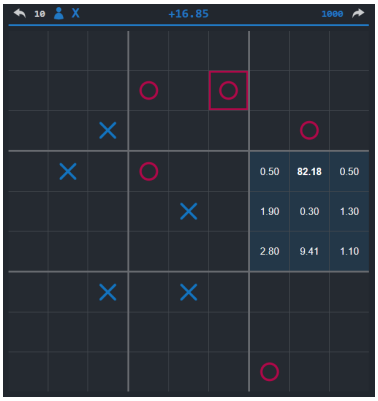
Figure 1: An example of tactical puzzle and its generated solution tree.

The generation of tactical puzzles for different difficulty levels was automated by conducting matches between agents of equal strength, with the search depth of both agents corresponding to the depth of the puzzle we wanted to find. We chose this approach to ensure that the resulting positions were interesting and balanced, as otherwise, the stronger side would usually have an overly obvious advantage at the start of the puzzle which would make it boring to solve.

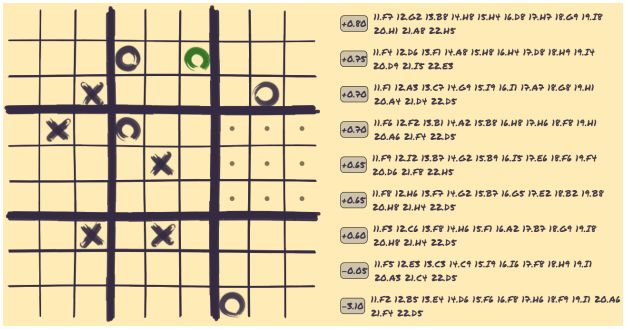
5.3 Strategic Puzzle Generation

Automating the creation of strategic puzzles is impossible without a program that could interpret the given position and simultaneously provide a human-understandable explanation. Additionally, generating strategic puzzles requires an agent with an advanced strategic understanding of the position, which our agents, using relatively simple heuristics, are incapable of. Therefore, we resorted to the most powerful freely available agent [13], which is based on the ideas of the AlphaZero program.

Thus, we generated the strategic puzzles manually. We searched for interesting and instructive positions that arose in games between the aforementioned agent and our stronger programs. We focused on moments when there was a significant deviation in the position evaluation between the two agents. When the agent with better strategic understanding detected an important change, we saved the given position, studied it more closely, and based on our understanding of the game, formulated a solution. The most common examples of such situations involved sacrificing the edge board to gain control over the central board. A basic example of this can be seen in Figure 2.



(a) User interface of the most powerful freely available agent. For the given position, it ran 1000 simulations and assessed the move F2 as the best with an 82% probability. It evaluates the position with a value of +16.85, which means it assigns approximately 58.4% win probability to player X (a value of 0 means a draw, 100 a win, and -100 a loss).



(b) Minimax agent with a search depth of 12. It marks the move F2 as the worst, as it does not recognize the long-term advantage.

Figure 2: Different interpretations of the same position, based on which we built the strategic puzzle.

6 Evaluation and Results

We conducted a quality analysis of the application with 14 volunteers. Their task was to use the app for an extended period to improve their knowledge of the game. We were interested in determining whether using the app had a positive impact on the development of their Ultimate Tic-Tac-Toe playing skills and

whether progress was dependent on motivation or the time spent learning.

To assess individual progress, participants played against the agent at the start of testing to determine their initial skill level. The application then tracked the highest level each user defeated, providing an estimate of their improvement over time. This progress, in relation to the number of puzzles solved, is illustrated in Figure 3. For a more concrete interpretation of obtained level strengths, refer to Section 4.

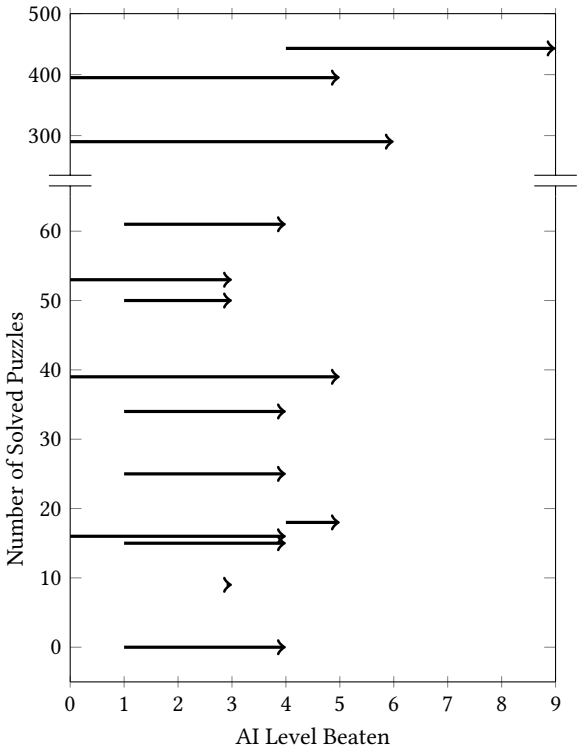


Figure 3: Progress in relation to the number of solved puzzles. Each arrow represents a human tester and indicates the change in the achieved level from the beginning to the end of the application's use.

7 Discussion

The results in Figure 3 indicate that solving more puzzles impacted users' ability to reach higher levels, but had less of an effect on lower levels. This is likely due to the fact that beginners can improve relatively quickly by simply playing the game, whereas advanced players require more effort to progress (eg. it is a lot easier to gain 100 rating points when you are rated 500 as compared to when you are rated 1500).

The reason for this is that at lower ratings, there is generally more room for rapid improvement because the skill gap between players tends to be more pronounced, and beginners can quickly benefit from fundamental knowledge and tactical awareness. As a result, achieving a higher rating initially is easier as players can fix obvious mistakes and exploit weaker opponents' errors.

However, as players reach higher levels, the competition becomes tougher, and the differences in skill become more nuanced. Players at this level are more consistent and less likely to make blunders, so improving further requires mastering advanced strategies, pattern recognition, and deeper positional understanding, making progress slower and more challenging. This reflects the diminishing returns on improvement as you climb the rating ladder.

It must also be mentioned that users were free to use any tools within the app during testing and solving more puzzles did not correlate with longer app usage. For a clearer assessment of puzzle significance, a controlled test focusing solely on puzzle-solving would be more appropriate.

8 Conclusion

In this work, we presented methods for generating puzzles for the game of Ultimate Tic-Tac-Toe. To evaluate the quality of these puzzles, we tracked how the number of solved puzzles impacted individual user progress. Our results indicate a correlation between the number of puzzles solved and the ability to reach stronger AI levels.

However, the evaluation could be refined by focusing exclusively on the puzzle-solving component, isolating it from other functionalities of the application. Additionally, the automation of tactical puzzle generation could be expanded to cover the middlegame phase, rather than being limited to endgame scenarios. Another area of improvement is providing clearer assessments of puzzle difficulty. This could be achieved by implementing a rating system that ranks puzzles based on completion rates, offering a more accurate measure of challenge for each puzzle.

Acknowledgements

The author would like to thank the family and friends who participated in testing the application.

References

[1] Guillaume Bertholon, Rémi Géraud-Stewart, Axel Kugelmann, Théo Lenoir, and David Naccache. 2020. At most 43 moves, at least 29: optimal strategies and bounds for ultimate tic-tac-toe. (2020). DOI: 10.48550/ARXIV.2006.02353.
[2] Chess.com. 2024. Chess.com. (June 2024). https://www.chess.com/.
[3] Rémi Coulom. 2006. Efficient selectivity and backup operators in monte-carlo tree search. In *In: Proceedings Computers and Games 2006*. Springer-Verlag.
[4] Justin Diamond. 2022. A practical method for preventing forced wins in ultimate tic-tac-toe. (2022). DOI: 10.48550/ARXIV.2207.06239.
[5] Nelson Elhage. 2020. Solving ultimate tic tac toe. (July 2020). https://minimax.dev/docs/ultimate/.
[6] Arpad E Elo. 1978. *The Rating of Chessplayers, Past and Present*. Arco Pub.
[7] Ofek Gila. 2019. Ultimate tic tac toe. (2019). https://www.theofekfoundation.org/games/UltimateTicTacToe/.
[8] Fernand Gobet and Peter J Jansen. 2006. Training in chess: a scientific approach. *Education and chess*.
[9] Henryk. 2023. Ultimate tic tac toe. (Dec. 2023). https://play.google.com/store/apps/details?id=com.henrykvdb.sttt.
[10] HPStudios. 2024. Ultimate tic tac toe. (June 2024). https://play.google.com/store/apps/details?id=com.MertTaylan.UltimateTicTacToe.
[11] Lichess. 2024. Lichess. (June 2024). https://lichess.org/.
[12] Levi Moore. 2020. Ultimate tic tac toe. (Nov. 2020). https://play.google.com/store/apps/details?id=com.ZeroRare.UltimateTicTacToe.
[13] Arkadiusz Nowaczynski. 2021. Ar-nowaczynski/utttai: alphazero-like ai solution for playing ultimate tic-tac-toe in the browser. (Dec. 2021). https://github.com/ar-nowaczynski/utttai.
[14] David Silver et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362, 6419, 1140–1144. eprint: https://www.science.org/doi/pdf/10.1126/science.aar6404. DOI: 10.1126/science.aar6404.
[15] Michael Xing. 2019. Ultimate tic-tac-toe. (Oct. 2019). https://www.michaelxing.com/UltimateTTT/v3/.

Ethical Consideration and Sociological Challenges in the Integration of Artificial Intelligence in Mental Health Services

Saša Poljak Lukek
sasa.poljaklukek@teof.uni-lj.si
University of Ljubljana, Faculty of Theology
Ljubljana, Slovenia

Abstract [1]
This article explores the transformative potential of artificial intelligence (AI) in the field of mental health, with a particular focus on ethical considerations and social challenges. As AI tools become increasingly sophisticated, their ability to support mental health interventions presents both opportunities and challenges. We discuss the importance of a human-centered approach to AI development and the need for comprehensive ethical guidelines to ensure patient safety and well-being. In addition, this paper explores key social trends such as the evolving dynamics of modern families, aging population, migration and considers how AI can be integrated into these contexts to improve mental health care.

Keywords:
Artificial Intelligence, Mental Health, Human-Centered Approach, Ethics, Modern Family Dynamics, Aging Populations, Migration

1 Introduction

1.1 Artificial intelligence in mental health services

Research on the application of AI in mental health care has shown some positive effects on the treatment of mental health problems [1], including early detection [2,3], providing feedback and personalized treatment plans [4], and developing of novel diagnose tools [2].

AI in mental health services is implemented through models like chatbots, digital platforms, and avatar therapy, enhancing accessibility and treatment options. Chatbots provide therapy via natural language processing [5], while digital platforms support online mostly cognitive behavioral therapeutic interventions [6]. Avatar therapy uses AI to help patients manage conditions like dementia, autism spectrum disorder, and schizophrenia [7].

1.2 The Prospect of artificial intelligence in mental health services

The future orientation underlines the importance of digital health in overcoming challenges such as limited access to services, especially in underserved regions, and outlines measures to ensure equitable access to digital health solutions across the European region [8]. The use of AI in mental health services raises questions about the role of non-human interventions, transparency in the use of algorithms and the long-term impact on the understanding of illness and the human condition [9]. There are also concerns about potential bias, gaps in ethical and legal frameworks, and the possibility of misuse [10,11].

However, there are at least two potentially positive effects of the use of AI in healthcare: Accessibility and personalization of services.

AI offers new mechanisms to reach those who might not otherwise be served. AI-supported tools can improve the early detection and diagnosis of mental disorders [12]. AI chatbots have shown promise in increasing referrals to mental health services, especially for minority groups who are blocked from accessing traditional care [13]. These technologies can provide initial assessments, psychoeducation and even treatment, expanding access to mental health support [12]. AI-driven virtual assistants and wearable devices enable continuous

[1] *This Publication is a Part of the Research Program The Intersection of Virtue, Experience, and Digital Culture: Ethical and Theological Insights, financed by the University of Ljubljana.*

monitoring and personalized care, which could improve patient outcomes [11,14].

The integration of artificial intelligence into mental health services represents a promising avenue for the development of personalized treatment plans through the sophisticated analysis of large datasets, enabling the identification of optimal therapeutic strategies tailored to specific client profiles [15,16]. This data-driven methodology enables the dynamic adaptation of therapy to the evolving needs of the client.

2 Overcoming Sociological Challenges through the Integration of Artificial Intelligence in Mental Health Services

2.1 Modern Family Dynamics

Modern family trends show that family structures and attitudes have changed significantly in recent decades [17]. There is a growing acceptance of different family forms, including unmarried cohabitation, same-sex relationships and joint custody arrangements [18]. These changes reflect an expansion of developmental idealism and increasing support for individual freedom in family choice [17].

On the other hand, there is a growing need for mental health services for families [19]. As the most vulnerable members of the family - the children - are usually also at risk, quick and effective action in family mental health is of great importance. Many families are struggling with various psychological problems. Together with the changing family structure, this means a great burden for every family member. In addition, access to psychologists, psychiatrics and therapists is limited, leading to an acute shortage of mental health professionals worldwide.

The accessibility of services is probably the strongest argument for the integration of AI in healthcare [12]. AI-powered conversational agents can improve the accessibility of mental health services by being available online at all times and in underserved areas, being scalable, reliable, fatigue-free, and providing consistent support, being culturally sensitive to adapt, and helping with education and symptom management.

2.2 Aging Populations

AI offers promising solutions for supporting an aging population, particularly in addressing cognitive decline and mental health challenges. AI applications can monitor vital signs, health indicators, and cognition, as well as provide support for daily activities [20]. With an increasing number of elderly individuals, AI can support mental health care by providing companionship through intelligent animal-like robots (e.g., Paro, Harp seal) and assisting in monitoring and managing conditions like dementia [21,22]. AI can also help in tracking cognitive health and providing timely interventions to maintain mental well-being in older adults. These technologies have the potential to enhance independent living and quality of life for older adults and their families.

2.3 Migration

Migrants often face mental health challenges due to displacement, cultural adjustment and language barriers. AI can help migrants access mental health services by providing culturally and linguistically relevant resources and support. Chatbots and AI-driven platforms can bridge gaps in care by providing immediate help and continuity of care across different regions [23].

Recent research highlights the increasing role of digitalization and artificial intelligence (AI) in migration and mobility systems, especially in the context of the COVID-19 pandemic [24]. While these technologies offer opportunities for improving human rights and supporting international development, they also bring challenges that require careful consideration of design, development and implementation aspects. The integration of AI into migration processes requires a focus on human rights at all stages that goes beyond technical feasibility and companies' claims of inclusivity [24].

3 Ethical Consideration in the Integration of Artificial Intelligence in Mental Health Services

One of the main caveats to the use of AI in mental health is the introduction of new ethical standards to ensure user safety. The approach to integrating AI into services should therefore be human-centered [25]. Any innovation should therefore focus on people in their most

vulnerable position. It is important to assess all risks with sufficient accuracy and avoid misuse of AI as much as possible. The most important areas for ethical consideration when integrating AI into mental health services should be privacy, bias, transparency, security.

Data privacy and security are critical in digital healthcare and require robust measures to protect sensitive information and prevent unauthorized access. Protecting privacy rights and ensuring informed consent are critical to maintaining trust and ethical standards in the use of personal health data [11]. Combining multiple data streams increases the risk of unauthorized use, which exacerbates privacy issues. Ensuring informed consent and maintaining transparency, especially in emergency operations, are critical to addressing these ethical concerns and protecting the rights of participants [26].

The use of AI in mental health treatment raises ethical concerns about bias, particularly among marginalized populations who are already discriminated against and lack access to mental health care. It is uncertain whether AI-assisted psychotherapy can effectively address cultural differences and close treatment gaps in diverse populations [27]. In addition, populations that are traditionally marginalized in fields such as psychology and psychiatry are most vulnerable to algorithmic biases in AI and machine learning [27,28]. These biases limit the ability of AI to provide culturally and linguistically appropriate mental health resources, exacerbating existing inequalities. The persistence of such biases in AI systems not only risks increasing health inequalities, but also exacerbates existing social inequalities and raises critical ethical considerations [9].

The future of artificial intelligence in clinical settings is affected by a significant ethical dilemma concerning the trade-off between the performance and interpretability of machine learning models [29]. The lack of transparency in AI models makes it difficult to detect and correct biases. This underscores the need for greater transparency to ensure ethical and fair clinical decision-making.

In summary, the integration of AI into mental health services requires the establishment of strict ethical standards to protect the safety and privacy of users. A human-centered approach is essential, with a focus on dealing with potential bias, especially among marginalized groups, the risks associated with data privacy and security, and the challenges posed by the lack of transparency of AI models.

4 Conclusion

We propose to define AI as a new ethical entity in the field of mental health [30]. AI represents a novel artifact that changes interactions, concepts, epistemic fields and normative requirements. This change requires a redefinition of the role of AI, which lies on a spectrum between a tool and an agent. This shift underscores the need for new ethical standards and guidelines that recognize the unique status of AI as a distinct and influential actor in the field of mental health.

The integration of AI into services can, on the one hand, provide more efficient and faster solutions to some of the sociological challenges of today's society, but on the other hand, requires a precise and correct definition of the limits within which these models can be used. These efforts aim to bridge the gap between technology and human-centered care and ensure that AI complements, rather than replaces, the therapeutic benefits of human interaction.

Literature

[1] Sandhya Bhatt. 2024. Digital Mental Health: Role of Artificial Intelligence in Psychotherapy. *Annals of Neurosciences*, 0,0, 1-11. doi:10.1177/09727531231221612

[2] Sijia Zhou, Jingping Zhao and Lulu Zhang. 2022. Application of Artificial Intelligence on Psychological Interventions and Diagnosis: An Overview. *Frontiers in Psychiatry*, 13(March), 1–7. https://doi.org/10.3389/fpsyt.2022.811665

[3] Klaudia Kister, Jakub Laskowski, Agata Makarewicz and Jakub Tarkowski. 2023. Application of artificial intelligence tools in diagnosis and treatmentof mental disorders. *Current Problems of Psychiatry*, 24, 1–18. https://doi.org/10.12923/2353-8627/2023-0001

[4] Rachel L. Horn and John R. Weisz. 2020. Can Artificial Intelligence Improve Psychotherapy Research and Practice? *Administration and Policy in Mental Health and Mental Health Services Research*, 47, 5, 852–855. https://doi.org/10.1007/s10488-020-01056-9

[5] Kerstin Denecke, Alaa Abd-alrazaq and Mowafa Househ. 2021. Artificial Intelligence for Chatbots in Mental Health: Opportunities and Challenges. In: Househ, M., Borycki, E., Kushniruk, A. (eds) *Multiple Perspectives on Artificial Intelligence in Healthcare*. 115–128. https://doi.org/10.1007/978-3-030-67303-1_10

[6] Elias Aboujaoude, Lina Gega, Michelle B. Parish and Donald M. Hilty. 2020. Editorial: Digital Interventions in Mental Health: Current Status and Future Directions. *Front. Psychiatry* 11, 111. doi: 10.3389/fpsyt.2020.00111

[7] Kay T. Pham, Amir Nabizadeh & Salih Selek. 2022. Artificial Intelligence and Chatbots in Psychiatry. *Psychiatric Quarterly*, 93, 1, 249–253. https://doi.org/10.1007/s11126-022-09973-8

[8] WHO. 2022. Regional digital health action plan for the WHO European Region 2023–2030 (RC72). (July 2022). Retrieved August 20, 2024 from https://www.who.int/europe/publications/i/item/EUR-RC72-5

[9] Amelia Fiske, Peter Henningsen and Alena Buyx. 2019. Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. *Journal of Medical Internet Research*, 21, 5, e13216. https://doi.org/10.2196/13216

[10] Elizabeth C. Stade, Shannon Wiltsey Stirman, Lyle Ungar, Cody L. Boland,H. Andrew Schwartz, David B. Yaden, Joao Sedoc, Robert J. DeRubeis, Robb Willer and Johannes C. Eichstaedt. 2024. Large Language Models Could Change the Future of Behavioral Healthcare: A Proposal for Responsible Development and Evaluation. *Mental Health Res* 3, 12. https://doi.org/10.1038/s44184-024-00056-z

[11] David B. Olawade, Ojima Z. Wada, Aderonke Odetayo, Aanuoluwapo Clement David-olawade, Fiyinfoluwa Asaolu and Judith Eberhardt. 2024. Enhancing mental health with Artificial Intelligence: Current trends and future prospects. *Journal of Medicine, Surgery, and Public Health*, 3, 100099. https://doi.org/10.1016/j.glmedi.2024.100099

[12] Koki Shimada. 2023. The Role of Artificial Intelligence in Mental Health: A Review. *Science Insights* 43, 5, 1119-1127. doi:10.15354/si.23.re820

[13] Max Rollwage, Johanna Habicht, Keno Juechems, Ben Carrington, Sruthi Viswanathan, Mona Stylianou, Tobias U. Hauser and Ross Harper. 2023. Using Conversational AI to Facilitate Mental Health Assessments and Improve Clinical Efficiency Within Psychotherapy Services: Real-World Observational Study. *JMIR AI*, 2, e44358. https://doi.org/10.2196/44358

[14] David D. Luxton. 2020. Ethical implications of conversational agents in global public health. *Bulletin of the World Health Organization*, 98, 4, 285–287. https://doi.org/10.2471/BLT.19.237636

[15] Leonard.Bickman. 2020. Improving Mental Health Services: A 50-Year Journey from Randomized Experiments to Artificial Intelligence and Precision Mental Health. *Adm Policy Ment Health*, 47, 795–843. https://doi.org/10.1007/s10488-020-01065-8

[16] Silvan Hornstein, Valerie Forman-Hoffman, Albert Nazander, Kristian Ranta and Kevin Hilbert. 2021. Predicting therapy outcome in a digital mental health intervention for depression and anxiety: A machine learning approach. *DIGITAL HEALTH*, 7, 1-11. doi:10.1177/20552076211060659

[17] Josef Ehmer. 2021. A historical perspective on family change in Europe. In Norbert F. Schneider and Michaela Kreyenfeld (eds). *Research Handbook on the Sociology of the Family,* 143–161. https://doi.org/10.4337/9781788975544.00018

[18] Keera Allendorf, Linda Young-Demarco and Arland Thornton. 2023. Developmental Idealism and a Half-Century of Family Attitude Trends in the United States. *Sociology of Development*, 9, 1, 1–32. https://doi.org/10.1525/sod.2022.0003

[19] WHO. 2022. World mental health report: transforming mental health for all. (June 2022) Retrieved August 20, 2024 from https://www.who.int/publications/i/item/9789240049338.

[20] Sara J.Czaja and Marco Ceruso. 2022. The Promise of Artificial Intelligence in Supporting an Aging Population. *Journal of Cognitive Engineering and Decision Making*, 16, 4, 182–193. https://doi.org/10.1177/15553434221129914

[21] Maria R. Lima. 2024. Home Integration of Conversational Robots to Enhance Ageing and Dementia Care. *ACM/IEEE International Conference on Human-Robot Interaction*, 115–117. https://doi.org/10.1145/3610978.3638378

[22] Wendy Moyle. 2019. The promise of technology in the future of dementia care. *Nature Reviews Neurology*, 15, 6, 353–359. https://doi.org/10.1038/s41582-019-0188-y

[23] Zahra Abtahi, Miriam Potocky, Zarin Eizadyar, Shanna L. Burke, Nicole M. Fava. 2022. Digital Interventions for the Mental Health and Well-Being of International Migrants: A Systematic Review. *Research on Social Work Practice*, 33, 5, 518-529. Doi:10.1177/10497315221118854

[24] Marie McAuliffe, Jenna Blower and Ana Beduschi. 2021. Digitalization and artificial intelligence in migration and mobility: Transnational implications of the covid-19 pandemic. *Societies*, 11, 4, 135. https://doi.org/10.3390/soc11040135

[25] Luke Balcombe and Diego de Leo. 2022. Human-Computer Interaction in Digital Mental Health. *Informatics*, 9,1, 14. https://doi.org/10.3390/informatics9010014

[26] Nicholas C. Jacobson and Matthew D. Nemesure. 2021. Using Artificial Intelligence to Predict Change in Depression and Anxiety Symptoms in a Digital Intervention: Evidence from a Transdiagnostic Randomized Controlled Trial. *Psychiatry Research*, 295, 113618. https://doi.org/10.1016/J.PSYCHRES.2020.113618

[27] Bennett Knox, Pierce Christoffersen, Kalista Leggitt, Zeia Woodruff and Matthew H. Haber. 2023. Justice, Vulnerable Populations, and the Use of Conversational AI in Psychotherapy. *American Journal of Bioethics*, 23, 5, 48–50. https://doi.org/10.1080/15265161.2023.2191040

[28] Zoha Khawaja and Jean C. Bélisle-Pipon. 2023. Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots. *Frontiers in Digital Health*, 5, 1278186. doi: 10.3389/fdgth.2023.1278186

[29] Danilo Bzdok and Andreas Meyer-Lindenberg,. 2018. Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3, 3, 223–230. https://doi.org/10.1016/J.BPSC.2017.11.007

[30] Jana Sedlakova and Manuel Trachsel. 2023. Conversational Artificial Intelligence in Psychotherapy: A New Therapeutic Tool or Agent? *American Journal of Bioethics*, 23, 5, 4–13. https://doi.org/10.1080/15265161.2022.2048739

Optimization Problem Inspector: A Tool for Analysis of Industrial Optimization Problems and Their Solutions

Tea Tušar
Jordan N. Cork
Andrejaana Andova
Bogdan Filipič
Jožef Stefan Institute and Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
{tea.tusar,jordan.cork,andrejaana.andova,bogdan.filipic}@ijs.si

Abstract

This paper presents the Optimization Problem Inspector (OPI) tool for assisting researchers and practitioners in analyzing industrial optimization problems and their solutions. OPI is a highly interactive web application requiring no programming knowledge to be used. It helps the users to better understand their problem by: 1) comparing the landscape features of the analyzed problem with those of some well-understood reference problems, and 2) visualizing the values of solution variables, objectives, constraints and any other user-specified solution parameters. The features of OPI are presented using a bi-objective pressure vessel design problem as an example.

Keywords

optimization, black-box problems, sampling, problem characterization, visualization

1 Introduction

Industrial optimization problems often require simulations to evaluate solutions. For example, in electrical motor design [18, 19], assessing the efficiency and electromagnetic performance of a proposed design is done by running a simulator that analyzes the motor magnetic field and flux distribution. Such evaluations are black boxes to the user and the optimization algorithm alike, i.e., the underlying functions cannot be explicitly expressed, which makes the problem hard to understand and solve.

The established way to gain a better understanding of industrial problems is through the analysis of their solutions. Depending on the problem at hand, this can be a challenging task, as industrial problems often have a large number of variables, multiple objectives and constraints [20].

The Optimization Problem Inspector (OPI) presented in this paper is a tool conceived to ease this task for both problem experts and optimization algorithm developers. OPI provides two ways to further the understanding of an optimization problem:

(1) It computes a set of landscape features of the analyzed problem and compares them to those of well-understood reference problems.
(2) It provides visualizations of solutions through the values of their variables, objectives, constraints and any other user-specified solution parameters.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.8265

OPI is a web application, implemented by a Python library called `optimization-problem-inspector` included in the PyPi Python package index[1]. It is highly interactive and requires no programming knowledge to be used.

Freely available contemporary software tools for multiobjective optimization, such as DESDEO [12], jMetal [7] (and jMetalPy [2]), the MOEA Framework [8], ParadisEO-MOEO [10], platEMO [17], pygmo [3], pymoo [4], and Scilab [15], provide the implementation of various optimization algorithms and test problems. While the majority of them include some visualization of solutions, the plots are mostly focused on showing algorithm results for the purpose of comparing algorithm performance and not to increase problem understanding. In addition, none of these tools compute additional problem features as OPI does. Therefore, OPI brings a unique perspective to optimization problem analysis and understanding.

Next, Section 2 presents the real-world problem that will be used to showcase the features of OPI in Section 3. The paper concludes with some remarks in Section 4.

2 Real-World Use Case

Our chosen real-world problem is a version of the well-known pressure vessel design problem, first proposed more than 30 years ago [16]. In this work, we adapt the formulation from [5] to handle the pressure vessel volume as a constraint, as well as an objective. We also remove one unnecessary constraint and use the original boundary constraints for the first two variables.

A pressure vessel is a tank, designed to store compressed gasses or liquids. It consists of a cylindrical middle part capped at both ends by hemispherical heads. The pressure vessel has four design variables (see Figure 1): the shell thickness, $x_1 = T_s$, the head thickness, $x_2 = T_h$, the inner radius, $x_3 = R$, and the length of the cylindrical section of the vessel, $x_4 = L$. The two thickness variables are integer multiples of 0.0625 inches, which correspond to the available thicknesses of rolled steel plates, while the length and the radius are continuous. The problem has three constraints, two on the search variables and one on

[1] https://pypi.org/project/optimization-problem-inspector/

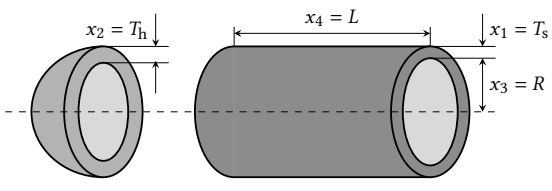


Figure 1: Pressure vessel design variables.

the volume. Its two objectives are to minimize the total costs, including the costs of the material, forming and welding, and to maximize the volume. The problem is formally defined as follows:

$$
\begin{aligned}
\min \quad & f_1(\mathbf{x}) = 0.6224 z_1 x_3 x_4 + 1.7781 z_2 x_3^2 + 3.1661 z_1^2 x_4 \\
& \quad + 19.84 z_1^2 x_3 \\
\max \quad & f_2(\mathbf{x}) = \pi x_3^2 x_4 + \frac{4}{3} \pi x_3^3 \\
\text{subject to} \quad & g_1(\mathbf{x}) = 0.0193 x_3 - z_1 \leq 0 \\
& g_2(\mathbf{x}) = 0.00954 x_3 - z_2 \leq 0 \\
& g_3(\mathbf{x}) = f_2(\mathbf{x}) \geq 1\,296\,000 \\
& x_1 \in \{18, \ldots, 32\} \\
& x_2 \in \{10, \ldots, 32\} \\
& x_3, x_4 \in [10, 200] \\
\text{where} \quad & z_1 = 0.0625 x_1 \\
& z_2 = 0.0625 x_2
\end{aligned}
$$

3 Optimization Problem Inspector Features

OPI is a web application, organized into five functional sections and a help section, providing guidance to the user. OPI expects the user to provide the problem specification and its data—evaluated problem solutions. Then, it generates and visualizes comparisons to artificial reference problems and visualizes the provided data.

Next, we will describe the main features of OPI through its five content sections: problem specification, sample generation, data, comparison to reference problems, and data visualization.

3.1 Problem Specification

In the first OPI section, the user can provide the specification of the industrial problem to be studied. The tool needs this information to properly generate the samples, described in the Section 3.2, and setup the visualisations.

The problem specification must be given in the `yaml` file format and needs to contain some basic information about problem parameters (variables, objectives, constraints) to be included in the analysis. OPI can handle one or more objectives and zero or more constraints. In addition to variables, objectives and constraints, the user can specify any number of other parameters that they want analyzed and visualized, for example, the name of the algorithm that found a solution or the time required to evaluate a solution.

For each of the parameters, the user needs to specify its name and its grouping (whether it is a variable, objective, constraint or something else). For variables, their type (continuous, integer or categorical) and the upper and lower bounds (for non-categorical types) are also required. An example `yaml` file, specifying a constrained multiobjective problem with several variables, is already provided within the tool to guide the user.

For the pressure vessel design problem, we can input four variables (first two are integer and last two are continuous), two objectives and three constraints. Alternatively, we can decide to skip the individual constraints and only use the total constraint violation instead.

3.2 Sample Generation

In OPI, a sample is a set of x-values, corresponding to the variables set in the problem specification section. In other words, a sample is a set of non-evaluated solutions.

If needed, the sample can be generated by the tool itself, based on the variable information provided in the problem specification step. However, this is not a required step in using OPI. A user that already has a set of (evaluated) solutions to work with can skip it and input the data directly (see Section 3.3).

Sample generation requires one to choose the number of desired samples, set to a default of 100, and the sample generation method. Three sample generation methods are supported: random, Sobol and Latin Hypercube, with random sampling being the default. The user may alter the settings of these sampling methods, such as the random generator seed. Selecting the button to generate and download the sample will download it in a `csv`-formatted file.

In the pressure vessel use case, OPI warns the user that not all sample generation methods are appropriate. In fact, the Sobol sampler and the Latin Hypercube Sampler are not compatible with non-continuous parameters. If used nevertheless, they may produce unexpected results.

3.3 Data

In OPI, the data is essentially a set of evaluated solutions, where each solution must contain a value for all objectives, constraints and other parameters included in the problem specification. The evaluation is conducted externally to the tool.

The data needs to be uploaded in a file in `csv` format. If any parameters from the problem specification are missing from the data, the tool will display a warning message. Any data parameters that are not included in the problem specification, are ignored without raising any warnings. When correctly input, the user will be able to view the data they have input, inspecting it in tabular format.

Inputting the data completes the setup stage of the process. The user may then begin generating visualisations to assist them in understanding their problem.

3.4 Comparison to Reference Problems

The first visualization mechanism provided by OPI visually compares the problem to a set of artificial reference problems with known properties. This is conducted by displaying the landscape features of the user-defined problem alongside the same features of each of the reference problems in a parallel coordinates plot. The plot is interactive—the user can highlight some of the problems by brushing along one of the parallel axes. In addition, the feature values can be viewed in a table and downloaded to a file in `csv` format.

The reference problems can be set by the user, however, confined within the collection labelled here as GBBOB, i.e., generalised BBOB, where BBOB stands for the well-known suite of 24 Black-Box Optimization Benchmarking problems with diverse properties [9]. OPI provides a generator of GBBOB problems that match the analyzed problem in terms of the number of variables and objectives and the presence or absence of constraints. For objectives and (optionally) the constraint, any single-objective BBOB problem instance can be used. The user can specify the desired GBBOB problems in the `yaml` format. OPI already contains five GBBOB problems to start.

A problem can be characterized by a large number of features, most hard to interpret by a human. In OPI, we included the following problem landscape features that are understandable to an expert user [1, 11, 13, 14]: `CorrObj`, `MinCV`, `FR`, `constr_obj_corr`, `H_MAX`, `UPO_N`, `PO_N` and a set of neighborhood features. `CorrObj`

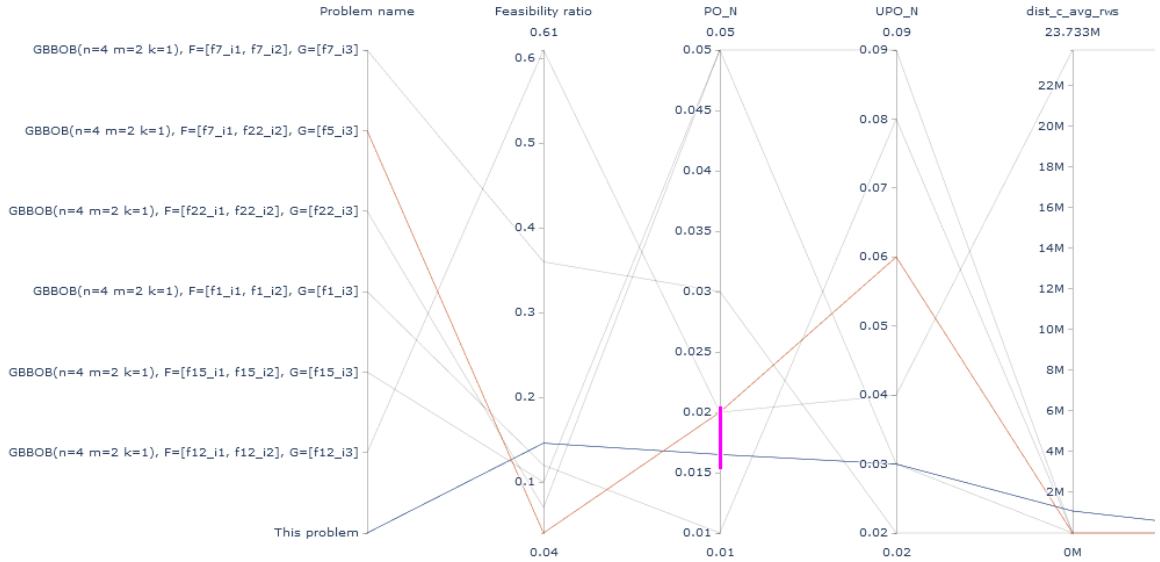


Figure 2: The initial part of the parallel coordinate plot visualizing feature values for the analyzed problem and the chosen set of artificial test problems.

is a feature that shows the correlation between the objectives. `MinCV` represents the minimum constraint violation among all solutions in the population. `FR` represents the proportion of feasible solutions in the population. `constr_obj_corr` presents the maximum correlation between the constraints and all the problem objectives. `H_MAX` is the maximum information content among all objectives. `UPO_N` is the proportion of unconstrained non-dominated solutions, while `PO_N` is the proportion of the constrained non-dominated solutions. The neighbourhood features denoted by `neighbourhood_feats` are a collection of features explaining the neighborhood of solutions, e.g., how many neighbors of a solution dominate the solution, how many neighbors are dominated by the solution, how many are incomparable to the solution, how close the neighboring solutions are, etc. OPI offers a total of 16 features, but the user can choose which to compute and visualize.

Figure 2 shows the initial part of the parallel coordinates plot (as the entire plot would not fit the paper) for the pressure vessel problem. In the comparison, we use the default five GBBOB reference problems as well as a custom created one. We notice that the pressure vessel problem is most similar to the custom GBBOB problem with the first objective equal to the step ellipsoid function f_7, the second to the multimodal peaks function f_{22}, and the linear constraint f_5. This similarity might be due to our mixed-integer problem containing plateaus in the continuous landscape space in which the features are computed, which is similar to the step ellipsoid function, and having linear constraints.

3.5 Data Visualization

In the data visualization section of the web application, the supplied data can be visualized using either a scatter plot matrix or a parallel plot. In both cases, the user can choose which problem parameters to visualize among all those listed in problem specification. Additionally, a simple data filtering that limits any variable between the desired minimum and maximum values is also supported and can be manipulated via the OPI interface in `yaml` format. The parameter used for coloring the solutions, as well as the color map, can also be specified by the user. Both visualizations support interaction and can be downloaded in `html` or `png` format.

3.5.1 Scatter Plot Matrix. The scatter plot matrix consists of n^2 plots for n chosen problem parameters as it contains 2-D scatter plots for all possible parameter pairs. In OPI, the user can apply brushing and linking to select the desired solutions in one or more of the scatter plots. These are then highlighted in all scatter plots in the matrix.

Figure 3 shows such a scatter plot for our pressure vessel problem. This visualization includes data from two sources. The first comes from a random sampling of the search space (shown in light blue) and the second from running the NSGA-II algorithm [6] on this problem for $2 \cdot 10^6$ function evaluations to achieve a good approximation of the Pareto front (shown in black). The two sources are set apart by a custom parameter that is then used for coloring the solutions. Some solutions from Figure 3 are highlighted – see the rectangle in the (x_3, x_1) scatter plot (third from the left in the top row).

These plots clearly show the linear relationship of the near-optimal solutions between x_1 and x_2 as well as x_1 and x_3. When only f_1 and f_2 are chosen, it is distinctively visible that the Pareto set approximation is piece-wise linear and disconnected.

3.5.2 Parallel Coordinates Plot. The parallel coordinates plot shows all chosen parameters as parallel coordinates and solutions as lines in the plot. Similarly as with the scatter plot matrix, interaction via brushing and linking is supported to select solutions that fit the desired values.

4 Conclusions

This work presented the features of Optimization Problem Inspector – a web application to support problem experts and algorithm designers in gaining a better understanding of industrial optimization problems. The tool provides comparisons to well-understood reference problems and interactive and highly-customizable visualizations, which can be exported in `html` and `png` formats.

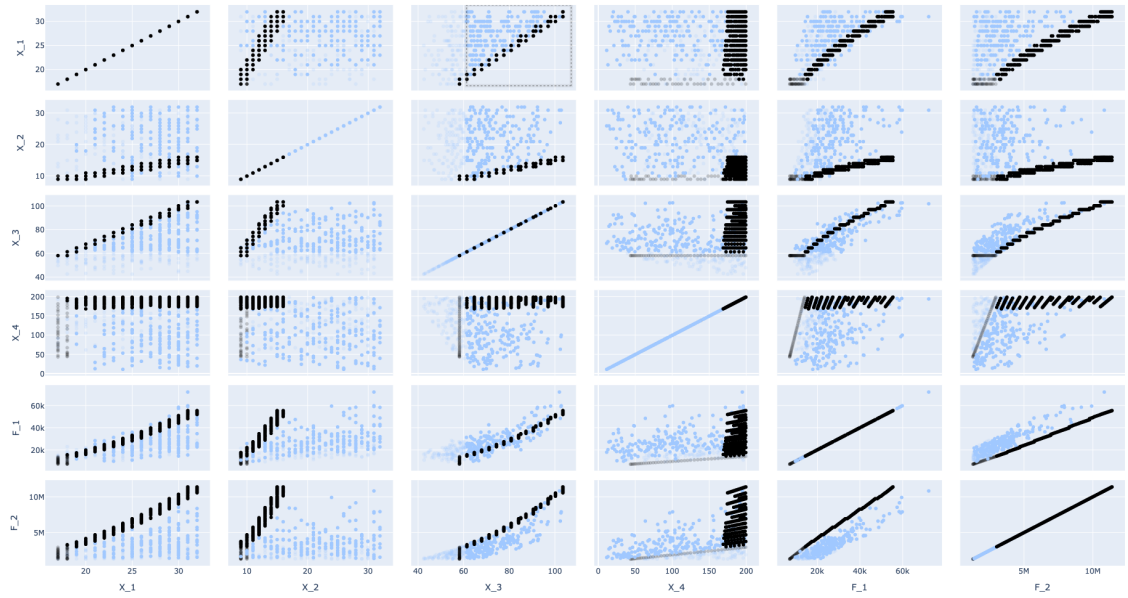


Figure 3: Random (light blue) and near-optimal (black) solutions of the pressure vessel design problem visualized in OPI with a scatter plot matrix containing variables x_1 to x_4 and objectives f_1 and f_2.

Samples can be exported and solutions imported using the standard `csv` format, which makes the data exchange between OPI and various optimization software easy to do. OPI functionality is made to be simple and at the same time flexible. Therefore, it is utilisable by non-experts and experts, alike, providing a wide range of angles from which to view the problems.

Acknowledgements

The authors acknowledge the financial support from the Slovenian Research and Innovation Agency (research core funding No. P2-0209 "Artificial Intelligence and Intelligent Systems", and project No. N2-0254 "Constrained Multiobjective Optimization Based on Problem Landscape Analysis") and from the Jožef Stefan Innovation Fund (project "A Tool for Analysis of Industrial Optimization Problems and Their Solutions"). This publication is also based upon work from COST Action "Randomised Optimisation Algorithms Research Network" (ROAR-NET), CA22137, supported by COST (European Cooperation in Science and Technology). We are grateful to Jernej Zupančič for implementing the core functionalities of the Optimization Problem Inspector.

References

[1] Hanan Alsouly, Michael Kirley, and Mario Andrés Muñoz. 2023. An instance space analysis of constrained multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation*, 27, 5, 1427–1439. DOI: 10.1109/TEVC.2022.3208595.

[2] Antonio Benítez-Hidalgo, Antonio J. Nebro, José García-Nieto, Izaskun Oregi, and Javier Del Ser. 2019. jMetalPy: A Python framework for multi-objective optimization with metaheuristics. *Swarm and Evolutionary Computation*, 51, 100598. DOI: 10.1016/J.SWEVO.2019.100598.

[3] Francesco Biscani and Dario Izzo. 2020. A parallel global multiobjective framework for optimization: pagmo. *Journal of Open Source Software*, 5, 53, 2338. DOI: 10.21105/joss.02338.

[4] Julian Blank and Kalyanmoy Deb. 2020. Pymoo: Multi-objective optimization in Python. *IEEE Access*, 8, 89497–89509. DOI: 10.1109/ACCESS.2020.2990567.

[5] Carlos A. Coello Coello. 2002. Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: A survey of the state of the art. *Computer Methods in Applied Mechanics and Engineering*, 191, 11, 1245–1287. DOI: https://doi.org/10.1016/S0045-7825(01)00323-1.

[6] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6, 2, 182–197. DOI: 10.1109/4235.996017.

[7] Juan José Durillo and Antonio J. Nebro. 2011. jMetal: A Java framework for multi-objective optimization. *Advanced Engineering Software*, 42, 10, 760–771. DOI: 10.1016/J.ADVENGSOFT.2011.05.014.

[8] David Hadka. 2024. MOEA Framework: A free and open source Java framework for multiobjective optimization. https://github.com/MOEAFramework/MOEAFramework. Computer software, version 4.4. (2024).

[9] Nikolaus Hansen, Steffen Finck, Raymond Ros, and Anne Auger. 2009. Real-Parameter Black-Box Optimization Benchmarking 2009: Noiseless Functions Definitions. Research Report RR-6829. INRIA. https://hal.inria.fr/inria-00362633v2.

[10] Arnaud Liefooghe, Laetitia Jourdan, and El-Ghazali Talbi. 2011. A software framework based on a conceptual unified model for evolutionary multiobjective optimization: ParadisEO-MOEO. *European Journal of Operational Research*, 209, 2, 104–112. DOI: 10.1016/J.EJOR.2010.07.023.

[11] K. M. Malan, J. F. Oberholzer, and A. P. Engelbrecht. 2015. Characterising constrained continuous optimisation problems. In *Proceedings of the 2015 IEEE Congress on Evolutionary Computation (CEC 2015)*, 1351–1358. DOI: 10.1109/CEC.2015.7257045.

[12] Giovanni Misitano, Bhupinder Singh Saini, Bekir Afsar, Babooshka Shavazipour, and Kaisa Miettinen. 2021. DESDEO: The modular and open source framework for interactive multiobjective optimization. *IEEE Access*, 9, 148277–148295. DOI: 10.1109/ACCESS.2021.3123825.

[13] Mario A. Muñoz, Michael Kirley, and Saman K. Halgamuge. 2015. Exploratory landscape analysis of continuous space optimization problems using information content. *IEEE Transactions on Evolutionary Computation*, 19, 1, 74–87. DOI: 10.1109/TEVC.2014.2302006.

[14] Cyril Picard and Jürg Schiffmann. 2021. Realistic constrained multiobjective optimization benchmark problems from design. *IEEE Transactions on Evolutionary Computation*, 25, 2, 234–246. DOI: 10.1109/TEVC.2020.3020046.

[15] Philippe Roux and Perrine Mathieu. 2016. Scilab: I. Fundamentals. In *Scilab, from theory to practice*. D-Booker Editions.

[16] E. Sandgren. 1990. Nonlinear integer and discrete programming in mechanical design optimization. *Journal of Mechanical Design*, 112, 2, 223–229.

[17] Ye Tian, Ran Cheng, Xingyi Zhang, and Yaochu Jin. 2017. PlatEMO: A MATLAB platform for evolutionary multi-objective optimization. *IEEE Computational Intelligence Magazine*, 12, 4, 73–87. DOI: 10.1109/MCI.2017.2742868.

[18] Tea Tušar, Peter Korošec, and Bogdan Filipič. 2023. A multi-step evaluation process in electric motor design. In *Slovenian Conference on Artificial Intelligence, Proceedings of the 26th International Multiconference Information Society (IS 2023)*. Vol. A. Jožef Stefan Institute, Ljubljana, Slovenia, 48–51.

[19] Tea Tušar, Peter Korošec, Gregor Papa, Bogdan Filipič, and Jurij Šilc. 2007. A comparative study of stochastic optimization methods in electric motor design. *Applied Intelligence*, 27, 2, 101–111. DOI: 10.1007/S10489-006-0022-2.

[20] Koen van der Blom, Timo M. Deist, Vanessa Volz, Mariapia Marchi, Yusuke Nojima, Boris Naujoks, Akira Oyama, and Tea Tušar. 2023. Identifying properties of real-world optimisation problems through a questionnaire. In *Many-Criteria Optimization and Decision Analysis: State-of-the-Art, Present Challenges, and Future Perspectives*. Natural Computing Series. Dimo Brockhoff, Michael Emmerich, Boris Naujoks, and Robin C. Purshouse, editors. Springer, 59–80. DOI: 10.1007/978-3-031-25263-1_3.

Multi-Agent System for Autonomous Table Football: A Winning Strategy

Marcel Založnik*
Jožef Stefan Institute
Ljubljana, Slovenia
marcel.zaloznik@gmail.com

Kristjan Šoln*
Faculty of Electrical Engineering, University of Ljubljana
Ljubljana, Slovenia
ks4835@student.uni-lj.si

Abstract

This paper presents a multi-agent system (MAS) for autonomous table football, developed for the FuzbAI competition at the University of Ljubljana. Our system consists of four independent agents, each dynamically assigned specific roles—Goalkeeper, Defender, Midfielder, and Attacker—based on real-time game analysis. This role-based architecture enabled seamless coordination between offensive and defensive strategies, allowing our team to secure first place. We describe the simulation framework used, the processing of sensor data, and the control strategies that allowed the agents to execute precise actions in a dynamic environment. The results highlight the effectiveness of adaptive, role-based decision-making, demonstrating the potential of MAS in real-time, competitive settings.

Keywords

multi-agent system, autonomous table football, role-based strategy, real-time decision making, AI in robotics

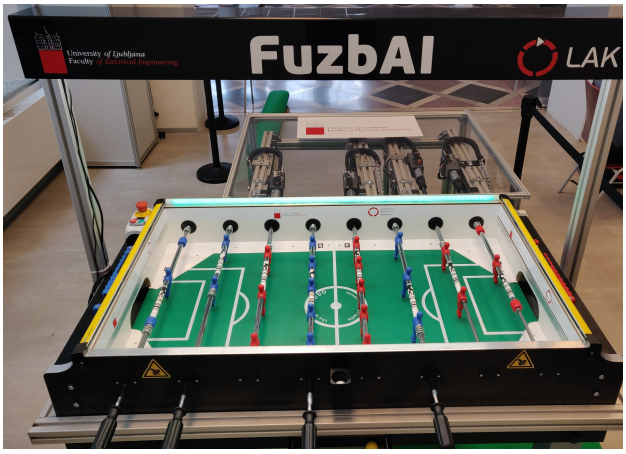


Figure 1: Table setup for the FuzbAI autonomous football competition.

1 Introduction

The FuzbAI competition, held as part of the "Dnevi Avtomatike" event at the Faculty of Electrical Engineering, University of Ljubljana, is a premier contest for students specializing in automation and artificial intelligence [11]. This event challenges participants to develop intelligent autonomous agents capable of playing table football without human intervention. The competition not only serves as a platform for demonstrating technical skills but also fosters innovation in the application of AI and machine learning techniques in real-time environments. Figure 1 illustrates the table setup used in the competition.

The FuzbAI competition is structured in a way that teams must design and implement a fully autonomous system capable of effectively competing against other AI-driven systems. Each match is a test of the participants' ability to integrate advanced algorithms and robotics, simulating the dynamics of a real football game on a miniature scale. The competitive format includes both qualification rounds and knockout stages, ensuring that only the most capable and innovative solutions advance to the final stages.

Our entry into the FuzbAI competition focused on the development of a multi-agent system (MAS), where each of our four rods functioned as an independent agent. These agents were designed to collaborate through a streamlined decision-making process, selecting roles that dictated their actions during gameplay. This strategic approach enabled our team to outperform competitors and ultimately secure first place in the competition.

This paper delves into the development and implementation of our multi-agent system. We will explore the architectural choices, the role-based decision-making strategies employed by each agent, and the overall system's performance in the context of the FuzbAI competition.

2 Competition Setup and System Description

The FuzbAI competition required all participants to develop programs capable of playing table football autonomously. To facilitate this, the competition provided a standardized simulation environment and a set of initial tools that every team used as the foundation for their development. This section describes the simulation framework, the types of data available from the system, and the means by which agents could interact with both the simulated and real game environments.

2.1 Simulation Framework

Participants were provided with a Python-based simulation framework designed to emulate a real table football match, as shown in figure 2. This simulator accurately replicated the physics of the game, including the movement of the ball and rods, and managed the interactions between the environment and the agents controlling the rods. The framework included fundamental functionalities such as ball tracking, rod positioning, and interaction rules, allowing all teams to concentrate on AI development without needing to construct the simulation infrastructure themselves.

One of the key features of the competition setup was that the interaction protocols for the simulator and the physical table were identical. The same signals and commands used to control

*Both authors contributed equally to this research.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.8341

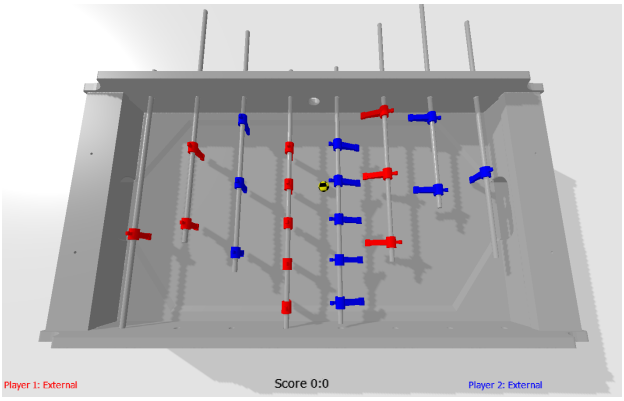


Figure 2: Simulator interface.

the actuators in the simulator were also used for the real table without any modification. This feature ensured that teams could seamlessly transition their algorithms from the simulated environment to the physical table setup, which was used in the final rounds of the competition. As a result, the simulation provided a consistent testing ground that mirrored the actual physical setup, enabling teams to develop and refine their strategies under uniform conditions.

2.2 Sensor Data

Both the simulation environment and the real table provided each team with data from two cameras, one placed on each side of the table. Each camera captured different views of the game, and teams had to decide how to combine the information from both cameras. The data provided by each camera included:

- Ball position: The coordinates of the ball on the 2D plane of the table.
- Ball speed: Velocity of the ball.
- Ball size: Area of the ball in the captured image (in pixels).
- Rod positions: Calibrated position of all rods (in the interval $[0, 1]$).
- Rod angles: Calibrated angle of all rods (in the interval $[-32, +32]$).

This camera data was streamed continuously, requiring teams to process and merge the inputs from both cameras to accurately interpret the game's state. The accuracy and frequency of the data were sufficient to enable real-time decision-making by the autonomous agents, whether interacting with the simulator or the physical table.

2.3 Actuator Outputs

To interact with the environment, each agent could send commands to the actuators that controlled the rods. The system allowed for two primary types of commands:

- Translatory movement: Moving the rod left or right across the table.
- Rotational movement: Rotating the rod to control the angle at which the players struck the ball.

Precise and timely commands were crucial for effective game control, as they enabled the agents to optimally position their figures, strike the ball accurately, and execute defensive or offensive strategies effectively.

3 Related Work

Research on multi-agent systems (MAS) and their application in robotic football has been extensively explored. This section reviews some contributions that have informed the development of autonomous systems for table football and real football.

Moos et al. (2024) [5] developed an automated football table as a research platform for reinforcement learning, highlighting the challenges of transferring learned behaviors from simulation to real-world environments and the need for robust algorithms to handle uncertainties. While reinforcement learning is a common approach in such studies, we did not achieve satisfactory results with it. Therefore, we decided to use multi-agent systems instead. Klančar et al. (2002) [4] investigated cooperative control in robot football (real football) using multi-agent systems, focusing on behavior-based control and dynamic role assignment among robots to optimize performance. Their approach emphasized effective communication for coordination in multi-agent settings. This work particularly inspired our approach to multi-agent systems, where we focused on behavior-based control and dynamic role assignment. Ribeiro et al. (2024) [6] proposed a probability-based strategy (PBS) for robotic football (real football), utilizing real-time data for centralized decision-making without relying heavily on pre-defined plays. Their approach demonstrated flexibility across different environments. Smit et al. (2023) [8] explored scaling multi-agent reinforcement learning (MARL) to a full 11v11 simulated football environment (real football), focusing on computational efficiency and the use of attention mechanisms to enhance scalability in large-scale multi-agent settings. Song et al. (2024) [9] conducted an empirical study on the Google Research Football platform (real football), introducing a population-based MARL training pipeline to quickly develop competitive AI players, highlighting the importance of scalable training frameworks. Scott et al. (2022) [7] examined end-to-end learning in RoboCup simulations (real football), optimizing both low-level skills and high-level strategies through competitive self-play, providing a comprehensive approach to multi-agent training in competitive environments.

4 MAS Approach to Autonomous Table Football Control

In this section, we describe the the methodology of our approach. We describe agent architecture, different agent roles and outline the actions they can take. Then, we discuss the conditions and priorities for role assignment during the game and evaluate the behavior of the system as a whole.

4.1 Agent Architecture

There exist several agent architectures, commonly used in MAS. Approaches, such as [4, 10, 12, 13], use role-based approach for interaction between agents and with the environment. In role-based approach, based on the concepts from role theory [1], the agents are assigned roles which affect their behavior. While the overall long-term goal of the system is typically predefined and does not change, e.g. win a table football match, the current role of an individual agent defines agent's short-term goals, which influences agent behavior, their decision-making process, and how they interact with the rest of the system. Furthermore, separation of agent functionality into independent roles can provide simplification and decoupling of individual tasks, leading to a more modular system, which can simplify and improve the extensibility of the implementation [3].

There exist several approaches to role and behavior implementation in MAS, such as merging different roles, role models and class members [2, 3, 4]. In our implementation, we simplify the architecture by allowing an agent to occupy only a single role at a time, and defining the roles in a way that allows reassigning between iterations of the algorithm without regard to the previous role or state of the agent.

Each role defines a set of possible actions an agent can take. The agents decide which action to take based on their priority and the current environment. More complex roles can be implemented in a stateful manner, meaning the decision on which action to take is dependent on previous actions as well. An agent can only be assigned a single role at a time, but can switch between roles throughout iterations regardless if the particular goal is fulfilled, when appropriate conditions arise. Additionally, every agent must have a role assigned at all times.

An action is a discrete, autonomous task that an agent can take on by making appropriate decisions and acting onto the environment, e.g. by sending commands to the actuators. This advances the agent toward the goal imposed by the current role. An agent can only execute a single action at a time. Additionally, every agent must be actively executing an action at all times.

These concepts were implemented using an Object Oriented approach, as suggested by the authors of the competition. In our implementation, each agent repeatedly executes a fast processing routine. Every iteration, the environment data is updated and role selection for the agent is performed. Then, as the agent decides on a role for that iteration, the appropriate role processing function is called, executing individual actions.

4.2 Role Description

A typical table football setup consists of four rods per player, each with a number of mounted figures. In this implementation, each rod is considered an agent, resulting in a system with four agents for which we define the following roles, typically associated with table football games.

Goalkeeper is the final line of defense, primarily responsible for intercepting the ball before it reaches the goal. Typically the left-most rod, which is nearest to the goal and has a single figure, the goalkeeper follows the ball position using two possible actions: *follow* and *misaligned follow*. The *follow* action simply tries to align the figure on the rod with the current ball position. However, if the velocity of the ball exceeds a predefined threshold, the agent instead attempts to estimate the ball trajectory based on its velocity vector. This estimation is simplified by assuming that the ball maintains a straight-line path. The figure is therefore positioned at the intersection of the rod and the estimated trajectory in an attempt to intercept a fast-moving ball.

The *misaligned follow* action is an augmented variant of the former action, designed to increase the overall defense surface of the defending agents. A common scenario in table football occurs when an attacker attempts to bypass the defenses by slightly pushing the ball parallel to the rod and striking it immediately after. Even though a human player might react fast enough to block such an attack, actuator response times are often insufficient. A defense strategy against such attacks is is to misalign the goalkeeper and defender figures, increasing the defense surface. Here, this is implemented by the *misaligned follow* action, and is activated whenever the ball is relatively slow, in the possession of the opponent and another agent in front of the Goalkeeper is currently in a Defender role. This decreases the chances of the opponent scoring even if the actuators fail to respond fast enough to block this style of attack. Here, communication between the two agents is performed implicitly, as each agent perceives the roles of other agents as a part of the overall environmental state.

Defender is an agent tasked with blocking opponent attacks by intercepting the ball when it is in the opponent's possession or moving towards the goal. This role utilizes a single *follow* action, similar to the Goalkeeper's *follow* action. Whenever the Defender role is active, the agent tracks the position and velocity of the ball, trying to match either its current coordinate or the estimated intersection with the trajectory of the ball. The agent identifies the figure closest to the intersection and attempts to move the rod using minimal amount of movement. This approach allows for faster adjustments during the game, improving defensive efficiency.

Midfielder is a an agent role with the primary task of raising the figures to allow passing the ball from behind the current agent. This role, although simple, is essential in order to avoid accidentally breaking a friendly attack by an Attacker agent behind the current rod.

Attacker is an agent with the task of kicking the ball towards the opponent goal in an attempt to score a point. Unlike other roles, the Attacker role is implemented in a stateful manner. Actions can only happen in a specified order, when the corresponding conditions are met. The role implements *follow*, *kick* and *prevent back-kick* actions.

Whenever the agent is assigned this role, the *follow* action is executed first. During the *follow* action, the agent slightly raises the figures in order to prepare for a kick. The figure closest to the ball is selected and rod offset is adjusted in order to align the figure with the ball. Whenever the agent determines that the alignment with the ball is sufficient, the agent moves onto the next state, the *kick* action. Here, the rod is rotated in order to strike the ball. During this state, it is still necessary to track the position of the ball, as the ball can move significantly within a few iterations of the algorithm. As the rod completes the forward rotation, the agent monitors the position of the ball and assesses if the figure successfully hit the ball. In that case, the next action is set back to *follow*, and the agent is usually assigned a new role according to the environment. However, if the figure missed the ball during the kick, the agent moves onto the *prevent back-kick* action. This final action is meant to prevent an accidental kick in the opposite of the intended direction. The rod is translated sideways and slowly rotated into a neutral position, in order to circumvent the ball. While executing this action, role switching for the current agent is disabled as well.

During execution, the agent aligns the rod position with the ball; however, a perfectly aligned figure results in a straight shot, which is easily defended by maintaining alignment with the ball. A more effective strategy involves kicking at an angle to aim for the goal or create a rebound off the wall, which is harder to defend. This role achieves this by slightly misaligning the figure before and during the kick. The agent computes the angle between the ball's current position and the selected target, with the figure's required misalignment set proportionally to this angle and adjusted by a tunable parameter for fine-tuning. This attack strategy significantly increases the performance of the Attacker role.

4.3 Role Assignment

Individual roles are assigned to agents according to defined assignment conditions and rules. Some approaches use an objective function in order to select a role, often taking role priority into account [4]. In this approach, we instead define a simple set of conditions which, along with role priority, decide on the most appropriate role for a particular agent based on the current state of the environment.

If in a particular instant, more roles fulfill the assignment conditions for a particular agent, the role with higher priority is selected. In this implementation, the highest priority belongs to the Attacker role, followed by the Goalkeeper, Defender and finally the Midfielder with the lowest priority. This ordering is based on the strictness of assignment conditions for each role, and the importance of that particular role. For example, the Attacker role has the strictest selection conditions among all roles, and therefore is assigned the highest priority, while the Midfielder role has a very broad assignment condition and is not as important compared to an Attack agent.

We define the role selection conditions as follows. The Attacker role is selected whenever the ball speed drops below a specified threshold, and the ball is within kicking clearance of the rod. The Goalkeeper role is selected if that particular agent belongs to the left-most rod, closest to the player's goal. The Defender role is selected whenever the ball is in front of the rod. Lastly, the Midfielder role is selected whenever the ball is behind the rod, as the role's only task is to raise the figures to allow the ball to pass forward.

This set of conditions combined with the defined role priority, allow the agents to switch between roles effectively and covers the main functionality required to play the game. Role priority ensures that the agent works toward a correct goal based on the circumstances. For example, any rod, even the Goalkeeper, should attempt to kick the ball if it is close and slow enough, while only the left-most rod should attempt to be the goalkeeper.

4.4 Behavior of the System as a Whole

The system's primary offensive strategy is for the Attacker agents to advance the ball as far forward as possible, ultimately aiming for the goal, while Midfielder agents ensure that they do not obstruct forward passes. During opponent attacks, the systems primary defensive strategy is for the Defender and Goalkeeper roles to intercept the ball. In certain situations, they collaborate to expand the defense surface, compensating for the limitations posed by actuator response times. Once the opponent's attack ends, agents detect the change in the environment and the roles are reassigned to shift the game towards offensive play.

The system's game strategy can be adjusted by modifying parameters such as role priority, assignment rules, or individual actions. For instance, a more defensive strategy can be achieved by tightening the conditions for assigning the Attacker role.

Overall, the implemented algorithm performs well, with the combination of discrete roles resulting in a competent gameplay. However, delays and noise present in measurements, and delays due to actuator response times, sometimes cause the system to miss, e.g. during attacks. The *prevent back-kick* action of the Attacker role proves essential in such situations, performing careful repositioning. Another surprisingly successful strategy is aiming at the goal or the wall during the *attack* action. Even if the ball does not follow the intended trajectory due to measurement noise and system delays, it still considerably increases the attack success rate. Additionally, even though there are no explicit, intentional passes between agents, the strategy of simply passing the ball as far forward as possible is enough for a successful gameplay.

The system overall is sensitive to changes in parameters and requires precise tuning. The simulator, although effective, does not perfectly simulate the physical table, and additional parameter tuning is required when transitioning from the simulator to real-world application.

5 Conclusion

This paper presented a multi-agent system (MAS) for autonomous table football, developed for the FuzbAI competition. Our role-based design allowed each rod to act as an independent agent, dynamically adapting to the game state. This approach enabled effective coordination between offense and defense, contributing to our system's first-place win.

The results demonstrate the effectiveness of a modular, adaptive architecture in dynamic environments, highlighting the importance of robust decision-making and quick role-switching. Future work could include machine learning to predict opponent behavior and optimize strategies, as well as expanding the system to more complex environments. Overall, our MAS showed strong performance in a competitive setting, offering valuable insights for future developments in autonomous systems.

References

[1] Bruce J Biddle. 1986. Recent developments in role theory. *Annual review of sociology*, 12, 1, 67–92.

[2] G. Cabri, L. Ferrari, and L. Leonardi. 2004. Agent role-based collaboration and coordination: a survey about existing approaches. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*. Vol. 6. IEEE. ISBN: 1062-922X. DOI: 10.1109/ICSMC.2004.1401064.

[3] E.A. Kendall. 1999. Role modelling for agent system analysis, design, and implementation. In *Proceedings - 1st International Symposium on Agent Systems and Applications and 3rd International Symposium on Mobile Agents, ASA/MA 1999*. IEEE, 204–218. ISBN: 0769503403. DOI: 10.1109/ASAMA.1999.805405.

[4] Gregor Klančar, Marko Lepetič, Boštjan Potočnik, Rihard Karba, and Drago Matko. Cooperative control of mobile agents in soccer game. Faculty of Electrical Engineering, University of Ljubljana, Slovenia, (2002).

[5] Janosch Moos, Cedric Derstroff, Niklas Schröder, and Debora Clever. 2024. Learning to play foosball: system and baselines. In Cornell University Library, arXiv.org. DOI: 10.48550/arxiv.2407.16606.

[6] António Fernando Alcântara Ribeiro, Ana Carolina Coelho Lopes, Tiago Alcântara Ribeiro, Nino Sancho Sampaio Martins Pereira, Gil Teixeira Lopes, and António Fernando Macedo Ribeiro. 2024. Probability-based strategy for a football multi-agent autonomous robot system. *Robotics*, 13, 1. DOI: 10.3390/robotics13010005.

[7] Atom Scott, Keisuke Fujii, and Masaki Onishi. 2022. How does ai play football? an analysis of rl and real-world football strategies. In *International Conference on Agents and Artificial Intelligence*. Vol. 1. Elsevier Scopus, 42–52. ISBN: 2184-3589. DOI: 10.5220/0010844300003116.

[8] Andries Smit, Herman A. Engelbrecht, Willie Brink, and Arnu Pretorius. 2023. Scaling multi-agent reinforcement learning to full 11 versus 11 simulated robotic football. *Autonomous agents and multi-agent systems*, 37, 1. DOI: 10.1007/s10458-023-09603-y.

[9] Yan Song, He Jiang, Zheng Tian, Haifeng Zhang, Yingping Zhang, Jiangcheng Zhu, Zonghong Dai, Weinan Zhang, and Jun Wang. 2024. An empirical study on google research football multi-agent scenarios. *International journal of automation and computing*, 21, 3, 549–570. DOI: 10.1007/s11633-023-1426-8.

[10] Manuela Veloso, Peter Stone, and Kwun Han. 1998. The cmunited-97 robotic soccer team: perception and multiagent control. In *Proceedings of the second international conference on Autonomous agents*, 78–85.

[11] Laboratorij za avtomatiko in kibernetiko. 2024. Dnevi avtomatike. Accessed: 2024-08-21. (2024). https://dnevi-avtomatike.si/?page_id=270.

[12] Franco Zambonelli, Nicholas R. Jennings, and Michael Wooldridge. 2003. Developing multiagent systems: the gaia methodology. *ACM transactions on software engineering and methodology*, 12, 3, 317–370. ObjectType-Article-2. DOI: 10.1145/958961.958963.

[13] Xiaoqin Zhang, Haiping Xu, and Bhavesh Shrestha. 2007. An integrated role-based approach for modeling, designing and implementing multi-agent systems. *Journal of the Brazilian Computer Society*, 13, 2, 45–60. DOI: 10.1007/bf03192409.

Towards a Decision Support System for Project Planning: Multi-Criteria Evaluation of Past Projects Success

Miha Hafner
Elea iC d.o.o.
Department for Tunnels and Geotechnics, and
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
miha.hafner@elea.si

Marko Bohanec
Jožef Stefan Institute
Department of Knowledge Technologies
Ljubljana, Slovenia
marko.bohanec@ijs.si

Abstract

Project planning typically refers to the project management step in which project assets, timelines, budgets, milestones, subcontractors, etc., are determined before the new project starts. In this paper, we address infrastructure design projects in the context of a specific company (Elea iC) and explore the idea of using data about past-finished projects to help project managers and project leaders in project planning. A crucial requirement in this context is the ability to evaluate/assess the success of finished/new projects. This paper proposes a solution using a multi-criteria model to evaluate finished projects. This way, we add project success information to the finished projects database, which we shall use in the decision support system being designed to extract knowledge for the new project plan.

Keywords

Project success evaluation, multi-criteria model, decision support systems, data analysis, data mining, project management, project leading tools.

1 Introduction

Infrastructure, such as tunnels, bridges, schools, houses, sewage systems, roads, etc., and its design discipline play a vital role in society. Thus, infrastructure design must have properly and thoroughly defined requirements, objectives, scope and constraints concerning many expert fields such as civil engineering, architecture, geology, geotechnics, environmental engineering, urban planning, and other expert fields [1], [2], [3]. The term design is connected to the process that ends with technical documentation, technical approvals, models, and other deliverables prepared at the end of the design process. Each such process is referred to as the project [4]. The projects are expected to have clearly defined:

- *Goals* defining the project's desired result, e.g., a building permit for a bridge, static analysis of a retaining wall, architectural design for a subway station, geotechnical exploration for a tunnel, etc. [4].
- *Objectives* that support project goals include concrete and measurable project characteristics such as deliverables, milestones, and other steps and strategies to achieve the goals [7].
- *Scope and requirements* concerning project boundaries, e.g., the need for experts, potential subcontractors, technical equipment and other requirements to finish the project.
- *Constraints and limitations* concerning project deadlines, costs, etc. [8].

Besides that, each project should finish with the client's and stakeholders' satisfaction [8].

To achieve the above for the new project, project planning is vital at the beginning of each new project [6], [8]. It is the project management and project leaders' task to recognize and include all these in the project plan so that the work and processes lead to successful project completion.

This study aims to support this process in the context of Elea iC company, an interdisciplinary provider of engineering services and projects in Slovenia [5]. We wanted to include the knowledge obtained from past–finished projects in the project planning process for the new projects. The company collected this data from 2001. The assumptions are as follows:

1. The finished projects in the database offer valuable information for the new project planning phase.
2. The project workflows established in the company and requirements remained similar over the years.

The main challenge related to this question is the new project success assessment and its consideration in light of the available finished project data [7]. Unfortunately, the actual finished projects database does not contain much information about the finished projects' success. To bridge this, we had to construct a project success evaluation model, evaluate finished projects in the database and add this information to the database. The expected result of those steps is a database suitable for applying data-analysis and knowledge extraction methods, such as hierarchical clustering and machine learning [20].

This paper describes the finished project success evaluation component (hereafter called FPSE), which is part of the future decision support system (DSS, [12], [13], [14]) for project planning (hereafter called E-DSS). First, we present the general architecture of the E-DSS, explaining the role and integration of FPSE in its context. In section 3, we present the database of finished projects and its preparation for supporting the configuration of new projects. The evaluation model used and the experimental evaluation of FPSE are presented in sections 4 and 5, respectively. Section 6 concludes the paper.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.8463

2 E-DSS Architecture

E-DSS is a DSS under construction to support the project management and project leaders in the Elea iC company (hereafter called "the user") in configuring the new project plan parameters when a new project starts.

The user is expected to define the E-DSS input as shown in Figure 1: the new project objectives, requirements, desires, and expectations. Practically, this means that the user collects all the available new project data by:

- Extracting the new project data from the new project assignment and contracts containing relevant information for the project planning.
- Checking the company and potential subcontractors' state of the resources and assets needed to complete the new project.

Examples of those data include projected monetary value, project scope and goals, project start and finish date, the expert fields needed for project completion, etc.

The E-DSS output (Figure 1) consists of the new project plan configuration together with the corresponding success scores (+S). The configuration comprises the data such as the number of employees involved, the number of subcontractors, work distribution, work duration, the number of pauses, etc. Project success scores are assessed assuming this configuration settings.

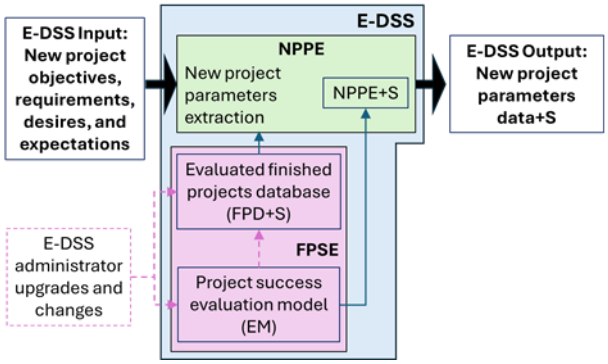


Figure 1: E-DSS architecture

Accordingly, E-DSS is composed of the following components (Figure 1):

- **NPPE** (New Project Parameters Extraction) is the component that extracts the potential new project configuration parameters and corresponding data to support the decision-making. NPPE is currently under construction and is aimed to operate interactively with the user and support: searching for similar projects in FPD+S according to a predefined range, searching the projects by desired success score, project segmentation, and project group identification—unsupervised descriptive analytics and parameter prediction by supervised machine learning methods. The component NPPE+S inside NPPE evaluates the success of the potential new project's configuration parameters obtained. The evaluation is made by EM, which is part of FPSE.
- **FPSE** (Finished Project Success Evaluation) consists of:
  - **EM** (Project success Evaluation Model) for evaluating the new project configuration (described in section 4).
  - **FPD+S** is the database of finished projects with project success evaluations (section 3).

Figure 1 also shows the element **E-DSS administrator** used to upgrade FPSE periodically by upgrading the database of the finished projects or making changes in EM according to the system's operational requirements and expected results.

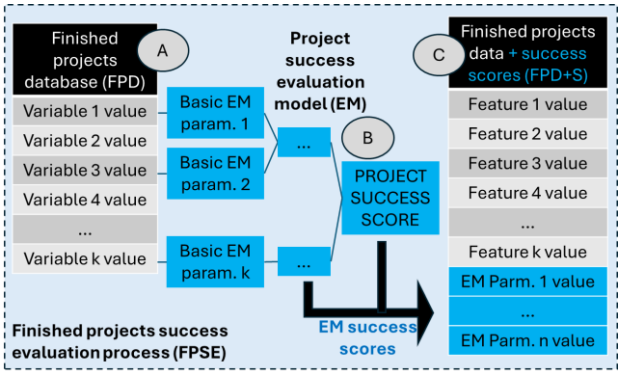


Figure 2: FPD+S development workflow

This paper focuses on the development of FPSE. The workflow is shown in Figure 2, consisting of the following steps:

Step A. Finished projects database preparation (FPD), as described in section 3.

Step B. Project success evaluation model (EM), as described in section 4.

Step C. Finished projects database with EM success scores (FPD+S): The result of the FPSE is the upgraded database of the finished projects with the finished projects' success scores (FPD+S).

3 Data Description

E-DSS is a data-driven system that operates on data from past-finished projects. This data was collected in Elea iC company from the year 2001 to 2023. At the beginning of the data collection, the number of the observed variables was relatively small, but it has grown substantially over the years. At the time of this study, the database contained data on 4704 finished projects, described by 39 numeric variables; 6 of them were date/time/year variables, and 2 categorical variables.

Data preparation (Step A, Figure 2) was carried out as follows:

1. Data cleaning: replacing "Nan" and deleting erroneous data;
2. Outlier's removal using the Interquartile range approach [18];
3. Data imputation: replacing the missing values using a descriptive statistic (e.g. mean, median, or most frequent) along each column or using a constant value [19]. We employed the mean strategy.
4. Sensitive data and information removal. For this reason, all numeric data was scaled to a range between 0 and 1.

We ended up with the database FPD containing data on 3132 finished projects described by 27 numeric variables. The variables describe the main project management characteristics, such as project financial results, workload distribution, number of employees, subcontractors, etc. Table 1 shows the list of all variables together with their basic statistics.

This way, the finished project database (FPD) was prepared for the FPSE component. FPD is the main resource for Exploratory Data Analysis for observing the data and its

properties, such as variable correlations, variable information gain, etc. These operations are invoked interactively by the user in the context of NPPE and are not discussed further in this paper.

Table 1: Basic statistics of the variables after data cleaning, outliers' removal, and data scaling

	min	max	median	mean
Project_Value	0.0	1.0	0.4215	0.2476
Number_of_Hours	0.0	1.0	0.0017	0.0123
Number_of_Employees_on_Project	0.0	1.0	0.0452	0.0420
Emploees_Workhours_Load	0.0	1.0	0.0119	0.0114
Costs_of_Subcontractors	0.0	1.0	0.0000	0.0069
Number_of_Subcontractors	0.0	1.0	0.0000	0.0150
Travelling_km	0.0	1.0	0.0000	0.0044
Project_Income	0.0	1.0	0.0018	0.0368
Project_Duration	0.0	1.0	0.2024	0.2414
Work_Period_Duration	0.0	1.0	0.0271	0.0770
Pauses_Duration	0.0	1.0	0.0000	0.0558
Number_of_Pauses	0.0	1.0	0.0000	0.0473
Proj_Work_Concentration	0.0	1.0	0.4985	0.5780
Number_of_Work_Months	0.0	1.0	0.0282	0.0740
Work_dist_Mean	0.0	1.0	0.0084	0.0190
Work_dist_Std	0.0	1.0	0.0283	0.0364
Work_dist_Min	0.0	1.0	0.0019	0.0076
Work_dist_25%	0.0	1.0	0.0041	0.0110
Work_dist_50%	0.0	1.0	0.0067	0.0161
Work_dist_75%	0.0	1.0	0.0096	0.0227
Work_dist_Max	0.0	1.0	0.0089	0.0253
Work_dist_Kurtosis	0.0	1.0	0.2051	0.2047
Work_dist_Skewness	0.0	1.0	0.4807	0.4799
Year	0.0	1.0	0.4167	0.4852
Pauses_Time_Share	0.0	1.0	0.0000	0.2096
Hour_Income	0.0	1.0	0.4547	0.2923
Time_Reserve	0.0	1.0	0.8128	0.6891

4 Evaluation of Projects' Success

The project success evaluation model (EM), developed in Step B (Figure 2), is aimed at:

- The evaluation of the projects in FPD resulting in the FPD+S (Figure 2).
- The evaluation of potential new projects suggested through interaction between the user and NPPE+S (Figure 1).

Project success evaluation involves multiple criteria that have to be aggregated into a single evaluation score. Different criteria might be of different importance and affect the score differently, i.e., with different weights. For this purpose, we chose MAUT (Multi-Attribute Utility Theory) [11], a multi-criteria modelling approach that facilitates both hierarchical structuring of criteria and using weights for the aggregation of scores.

Considering the above requirements, available FPD data and other multi-criteria approaches to project evaluation ([15], [16], [17]), we developed the EM as presented in Figure 3.

EM consists of three components [10]: input parameters, evaluation parameters and aggregation functions.

Input parameters are variables in the leaf nodes of the model:

- *Project Work Concentration*: explains the distribution of the work on the project. If the value is closer to 0 or 1, the majority of the work is done at the beginning or at the end of the project, respectively.
- *Time Reserve*: explains if the project work ended earlier than defined in the contract.
- *Number of Pauses*: the number of times the work on the project stopped.
- *Pauses Time Share*: the ratio between the months the employees did not work and the total number of months.
- *Hour Income*: the ratio between project value and the number of work hours necessary to finish the project.

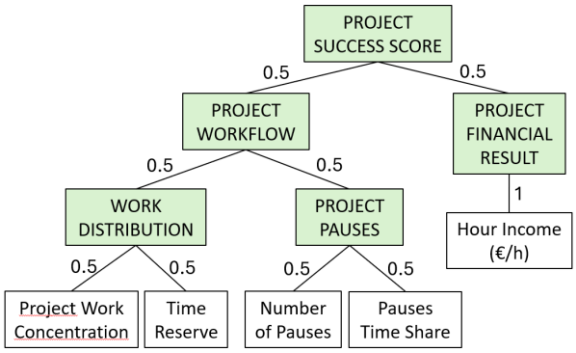


Figure 3: Multi-criteria model for the projects' success evaluation

Evaluation parameters represent outputs of the model*:*

- *WORK DISTRIBUTION*: assesses the characteristics of the work distribution in the project duration.
- *PROJECT PAUSES*: assesses the work pauses.
- *PROJECT WORKFLOW*: combines evaluation parameters WORK DISTRIBUTION and PROJECT PAUSES
- PROJECT FINANCIAL RESULT: assesses the project's success from the financial point of view.
- *PROJECT SUCCESS SCORE*: overall success score, determined by aggregation of all subordinate parameters.

Aggregation functions map subordinate EM parameters to the corresponding parent parameters. Employed is the weighted average function, using weights shown in Figure 3. Currently, weights are chosen to make all parameters equally important.

5 Experimental Evaluation of FPSE

Figure 4 shows an example of evaluating a project from FPD. Input parameters' values (terminal nodes) were obtained from the data base, while evaluation parameters' values (green nodes) were calculated by EM. The example project shows good workflow score (0.75), but has a poor financial score (0.29), both leading to an average success score (0.52) of the project. Several other projects of different types were evaluated in this way, confirming the appropriateness of EM structure and conformance with requirements of potential users. In this way, the quality of EM was assessed on a sample of past projects. Further assessment is planned in the next stages while configuring new projects, where EM's results can be confronted with opinions of project leaders actively involved in the process.

EM already enables evaluation of multiple finished projects. In Step C (Figure 2), FPD was extended by adding five variables corresponding the five Evaluation parameters of EM. All projects in FPD were evaluated by EM, resulting in FPD+S.

Basic statistics of FPD+S is presented by the distribution of the variables in Figure 5. The variables marked with red colour on the x-axis are E-DSS input parameters, the green uppercase variables are those corresponding to success scores, and the blue variables are potential new project parameters. The distribution of the final project evaluation, PROJECT_SUCCESS_SCORE,

(average = 0.52, min = 0.15, max = 0.94) indicates that it well covers the range of possible outcomes and enables the discrimination and sorting of projects.

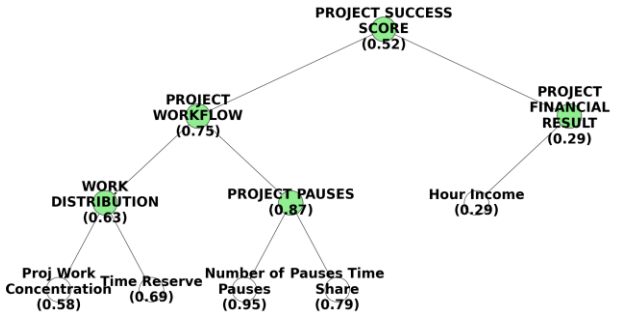


Figure 4: Example of evaluating a project using EM

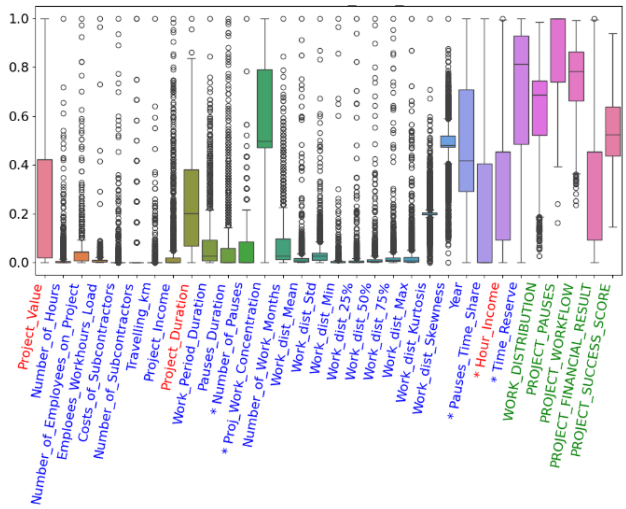


Figure 5: Distribution of the FPD+S features, including EM project success assessments

6 Conclusions

E-DSS is a DSS under construction, aimed at supporting the project management and project leaders' process in the new infrastructure project planning phase. We presented the design and development of the FPSE component, consisting of a multi-criteria project success evaluation model EM and a data base of projects, extended with success evaluation scores FPD+S.

EM has been developed using the MAUT approach and has turned out to be "fit-for-purpose". It employs the data that is available in the projects' database. It meaningfully describes aspects of the project's success and offers practical and functional model for the evaluation of multiple projects in the database.

FPSE is a key decision-support resource for E-DSS. E-DSS will allow the user to interactively search for similar past projects, to filter them according to the success score and simulate the effects of alternative project configurations, ultimately proposing the best one. Approaches based on unsupervised descriptive analytics (clustering) and supervised machine learning methods for prediction of E-DSS output parameters are foreseen for this purpose. Actually, we have already tested hierarchical clustering and decision tree classification methods on FPD+S, and first results are encouraging. We obtained meaningful clusters of past projects and created decision trees for prediction of individual output parameters that may lead to high new project success scores.

Future work will primarily continue by further data analysis and data mining of FPD+S, attempting to design effective algorithms for interactive exploration of past projects and suggesting as good as possible configurations of new projects. On this basis, we shall make a detailed functional specification of the NPPE+S component and design/implement the E-DSS.

Despite that E-DSS considered here is tailor-made for the specific business environment and is bound to the specific data base, the approach seems general enough to be applied to similar environments, projects and processes [9]. This work is a showcase of substantial efforts needed to prepare a corporate database for decision-support, which is often neglected in the literature. The main contribution is a combination of data processing with MAUT-based multi-criteria decision modelling.

References

[1] Fransje L. Hooimeijer, Jeremy D. Bricker, Adam J. Pel, A. D. Brand, Frans H.M. Van de Ven, Amin Askarinejad. 2022. Multi-and interdisciplinary design of urban infrastructure development. In Proceedings of the Institution of Civil Engineers: Urban Design and Planning. Vol.175. TU Delft. 153-168.

[2] Simon Christian Becker, Philip Sander. 2023. Development of a Project Objective and Requirement System (PORS) for major infrastructure projects to align the interests of all the stakeholders. In Expanding Underground - Knowledge and Passion to Make a Positive Impact on the World. CRC Press, London, UK, 3369-3376. DOI:10.1201/9781003348030-408.

[3] Michel-Alexandre Cardin, Ana Mijic, Jennifer Whyte. 2023. Data-driven infrastructure systems design for uncertainty, sustainability, and resilience. In D. M. Fabio Biondini, Life-Cycle of Structures and Infrastructure Systems. CRC Press, London, UK, 2565 – 2572. DOI: 10.1201/9781003323020-312.

[4] Saša Žagar. 2016. Organizacijski model v projektivnem podjetju Elea iC d.o.o.. Maribor, B.Sc. Thesis, Retrieved July 12, 2024 from https://dk.um.si/IzpisGradiva.php?id=58799&lang=eng.

[5] Elea iC webpage. https://www.elea.si/en/.

[6] Jürg Kuster, Eugen Huber, Robert Lippmann, Alphons Schmid, Emil Schneider, Urs Witschi, Roger Wüst. 2015. Project Management Handbook. Springer-Verlag, Berlin Heidelberg, Germany.

[7] Anton Hauc. 2007. Projektni management. (2nd. ed.). GV Založba, Ljubljana, Slovenija.

[8] Harvey A. Levine. 2002. Practical Project Management: Tips, Tactics, and Tools. John Wiley & Sons, Inc., New York, NY.

[9] Nadja Damij, Talib Damij. 2014. Process management. Springer-Verlag, Berlin Heidelberg, Germany.

[10] Marko Bohanec. 2012. Odločanje in modeli. DMFA – založništvo, Ljubljana, Slovenija.

[11] Salvatore Greco, Matthias Ehrgott, José Rui Figueira. 2016. Multiple Criteria Decision Analysis, State of the Art Surveys. Springer, Portsmouth, UK. DOI 10.1007/978-1-4939-3094-4

[12] Maria Rashidi, Maryam Ghodrat, Bijan Samali and Masoud Mohammadi. 2018. Decision Support Systems. In Management of Information Systems. IntechOpen, London, UK, 19-38. DOI: 10.5772/intechopen.79390.

[13] Daniel Joseph Power. 2013. Decision Support, Analytics, and Business Intelligence. Business Expert Press, New York, NY. DOI 10.4128/9781606496190.

[14] Sofiat Abioye, Lukumon Oyedele, Lukman Akanbi, Anuoluwapo Ajayi, Juan Manuel Davila Delgado, Muhammad Bilal, Olugbenga Akinade, Ashraf Ahmed. 2021. Artificial intelligence in the construction industry: A review of present status, opportunities and future challenges. Journal of Building Engineering 44, Elsevier. https://doi.org/10.1016/j.jobe.2021.103299.

[15] Erwin Berghuis. 2018. Measuring Systems Engineering and Project Success, Master's Thesis. University of Twente. https://purl.utwente.nl/essays/75088

[16] Ali Beiki Ashkezari, Mahsa Zokaee, Amir Aghsami, Fariborz Jolai, Maziar Yazdani. 2022. Selecting an Appropriate Configuration in a Construction Project Using a Hybrid Multiple Attribute Decision Making and Failure Analysis Methods. Buildings, MDPI, Volume 12, 643. DOI: https://doi.org/10.3390/buildings9050112.

[17] Urban Pinter, Igor Pšunder. 2013. Evaluating construction project success with use of the M-TOPSIS method. Journal of civil engineering and management, Volume 19(1), 16–23. doi:10.3846/13923730.2012.734849

[18] Interquartile range. Retrieved May 15, 2024 from https://en.wikipedia.org/wiki/Interquartile_range

[19] SimpleImputer. Retrieved May 15, 2024 from https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html

[20] Aggarwal C. Aggarwal. 2015. Data Mining: The Textbook. Springer, New York, USA.

Minimizing Costs and Risks in Demand Response Optimization: Insights from Initial Experiments

Mila Nedić
Faculty of Mathematics and Physics
University of Ljubljana
Ljubljana, Slovenia
mn38120@student.uni-lj.si

Tea Tušar
Jožef Stefan Institute and
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
tea.tusar@ijs.si

Abstract

This paper presents a method for changing the energy use of consumers participating in Demand Response (DR) programs, focusing on peak balancing to improve grid stability. Multiple objectives including costs and risks are considered, and a weighted sum is used to transform them into a single objective. This results in an optimization problem that can be optimally solved. To calculate the costs, the load consumption baseline needs to be established. Since this is challenging and can be exploited, we conduct initial experiments to test whether our method to adjust the baseline can be easily manipulated. We explore an original scenario and three of its variants to examine the effects of various parameters on the optimization outcome. Our results indicate that 1) an excessive emphasis on risk results in no energy change, 2) enforcing a net zero energy change minimizes energy use while still securing the rebate, and 3) without an adjustment period, the consumer is less inclined to increase the load just before the demand period. In future work, we will reformulate some objectives to avoid exploitation and better reflect the real-world needs of DR.

Keywords

multiobjective optimization, mixed-integer linear programming, demand response, baseline consumption, electrical grid

1 Introduction

Peaks in energy demand can strain the electrical grid, leading to inefficiencies and potential failures. A widely used strategy for balancing these peaks is Demand Response (DR), in which the Distribution System Operator (DSO) forecasts future peaks and requests from consumers to adjust their energy use to reduce them. In the peak time rebate DR program [2], consumers receive a rebate if they reduce their load in the demand period. On the other hand, if they commit to respond to the demand, but fail to do so, they can be penalized. It is therefore of utmost importance to accurately assess whether and how much a consumer reduced their load to meet the demand.

The load reduction of a consumer is computed as the difference between its baseline (the amount of energy the customer would have consumed without a demand request) and its actual use [2]. The importance of establishing a baseline and the various ways of calculating it are presented in [5]. Common methods for calculating baselines include simple historical data averages, exponential moving averages and short-term load forecasting techniques. However, baselines can be exploited, e.g., when consumers artificially increase consumption before an event to inflate their baseline and maximize the awarded rebate.

Through the SEEDS project[1], we are developing a methodology for providing energy flexibility services to prosumers – participants in energy markets capable of both producing and consuming energy – in order to enhance grid stability. Machine learning is used to predict the baseline energy usage of prosumers and their flexibility, while mixed-integer linear programming (MILP) is used to optimize the operation of prosumers within their flexibility. Our approach will be tested in the Slovenian pilot, in collaboration with Petrol d.d. and Elektro Celje d.d.

Our work integrates prosumer flexibility into DR optimization, focusing on minimizing costs and risks while limiting energy fluctuations. While the goal is to eventually use this approach on real-world data from the pilot, this paper reports on some initial experiments verifying whether the current problem formulation results in solutions with desired properties. In particular, we wish to test if our adjusted consumer baseline approach can be easily exploited.

Research on prosumer flexibility, optimization techniques, and demand response optimization includes a wide range of approaches [8]. In [3], Balázs et al. quantify residential prosumer flexibility using engineering models and real-world data. Their work provides valuable insight into prosumer behavior and energy management. Capone et al. [4] optimize district energy systems by balancing costs and carbon emissions with genetic algorithms and linear programming, showing significant emission reductions at a modest cost increase. Magalhães and Antunes [7] compare thermal load models in demand response strategies using MILP, finding that discrete control formulations improve computational efficiency. Thus, our methodology is in line with related work while the actual optimization problem (its variables, objectives and constraints) differs from existing ones as it is adapted to our specific use case.

This paper is further organized as follows. In Section 2, we provide a brief overview of the optimization problem, followed by its detailed definition in terms of its variables, constraints and objectives. The optimization approach is explained in Section 3, where we discuss the scalarization technique used to transform our multi-objective problem into a single-objective MILP form and the method used to solve it. The experiments and their results are given in Section 4. Finally, conclusions and further work ideas are described in Section 5.

2 Optimization Problem

The problem formulation in this work assumes a peak time rebate DR program in which the DSO and the consumer have a contract stipulating the following conditions: 1) the consumer can chose whether to respond to a demand request, 2) if the consumer

[1] https://project-seeds.eu/

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.8587

participates in DR, it receives a rebate proportional to the reduced load, 3) if the consumer participates in DR but does not reduce the load by at least 75 % of the required amount, it is penalized, 4) the load reduction is estimated using an adjusted consumer baseline, which takes into account the forecast consumer energy usage as well as its actual consumption before the demand period.

The optimization task is to set the energy consumption of all loads of a consumer participating in DR taking into account their flexibility so that consumer costs, risks and energy fluctuations are minimized. This ensures efficient grid operation while maintaining economic feasibility for the consumer.

To formally define our optimization problem, we first introduce its variables, followed by the constraints and the objective functions we aim to optimize. Finally, we provide an overview of the weighted sum approach, which serves as the scalarization technique to transform all objective values into a single one.

2.1 Variables

A solution is specified by the energy amounts $E_{c,i} \in \mathbb{R}$ for each consumer load $c \in C$ and time interval $i \in \{1, \ldots, n\}$. They correspond to the change of consumption from the forecast one. These are the only variables of this optimization problem.

From these energy amounts and the forecast timetable of energy usage, the resulting energy consumption E_i in time interval $i \in \{1, \ldots, n\}$ is computed as

$$E_i = E_i^{\mathrm{F}} + \sum_{c \in C} E_{c,i}.$$

2.2 Constraints

The energy amounts of a solution need to adhere to two kinds of constraints. The first type are the interval energy constraints:

$$E_{c,i}^{\min} \leq E_{c,i} \leq E_{c,i}^{\max},$$

for each consumer load $c \in C$ and time interval $i \in \{1, \ldots, n\}$. The second are the total energy constraints:

$$E_c^{T,\min} \leq \sum_{i=1}^{n} E_{c,i} \leq E_c^{T,\max},$$

for each consumer load $c \in C$.

2.3 Objective Functions

The three objectives to be minimized in this scenario are the costs, risks and energy fluctuations.

The first optimization objective f_1 consists of all costs associated with the solution and equals

$$f_1 = f^{\mathrm{E}} - f^{\mathrm{R}} + f^{\mathrm{P}},$$

where f^{E} represents the energy costs, f^{R} is the rebate for the recognized load reduction and f^{P} is the penalty that is charged in case the recognized load reduction does not meet the requirements.

The energy costs f^{E} equal the sum of energy costs over all time intervals $i \in \{1, \ldots, n\}$,

$$f^{\mathrm{E}} = \sum_{i=1}^{n} p_i \, E_i,$$

where p_i is the interval energy price.

The solution gains a rebate it the load is reduced in the demand period $\{t^{\mathrm{S}}, \ldots, t^{\mathrm{E}}\}$. Note that the recognized load reduction E_t^{R}, $t \in \{t^{\mathrm{S}}, \ldots, t^{\mathrm{E}}\}$, is computed from the adjusted timetable energy E_t^{A} instead of the forecast one E_t^{F}, where the adjustment is determined by the energy amounts in the adjustment period – the n^{A} intervals before the start of the demand period t^{S}. More formally, the adjusted timetable is computed as

$$E_t^{\mathrm{A}} = \begin{cases} E_t^{\mathrm{F}} - \dfrac{1}{n^{\mathrm{A}}} \displaystyle\sum_{j=t^{\mathrm{S}}-n^{\mathrm{A}}}^{t^{\mathrm{S}}-1} \left(E_j^{\mathrm{F}} - E_j\right), & \text{if } n^{\mathrm{A}} > 0; \\ E_t^{\mathrm{F}}, & \text{otherwise} \end{cases}$$

for all intervals $t \in \{t^{\mathrm{S}}, \ldots, t^{\mathrm{E}}\}$ in the demand period. Then, the recognized load reduction E_t^{R} at demand time interval $t \in \{t^{\mathrm{S}}, \ldots, t^{\mathrm{E}}\}$ is determined as

$$E_t^{\mathrm{R}} = E_t - E_t^{\mathrm{A}},$$

while the total recognized load reduction E^{R} is computed as

$$E^{\mathrm{R}} = \sum_{t=t^{\mathrm{S}}}^{t^{\mathrm{E}}} E_t^{\mathrm{R}}.$$

A rebate is awarded if E^{R} is negative (the consumption has been reduced). If the total recognized load reduction exceeds the total demanded energy reduction E^{T}, the rebate is capped, i.e.,

$$f^{\mathrm{R}} = \begin{cases} p^{\mathrm{B}} \min\left(\left|E^{\mathrm{R}}\right|, \left|E^{\mathrm{T}}\right|\right), & \text{if } E^{\mathrm{R}} < 0 \\ 0, & \text{otherwise} \end{cases}.$$

Finally, a penalty is added to the total costs if the demand has not been met, that is, the ratio between the recognized and demanded energy reduction, E^{D}, in any of the demand time intervals $t \in \{t^{\mathrm{S}}, \ldots, t^{\mathrm{E}}\}$ is lower than 75 %,

$$f^{\mathrm{P}} = \begin{cases} p^{\mathrm{P}} \left|E^{\mathrm{T}}\right|, & \text{if } \dfrac{E_t^{\mathrm{R}}}{E^{\mathrm{D}}} < 75\,\% \text{ for one or more } t \in \{t^{\mathrm{S}}, \ldots, t^{\mathrm{E}}\} \\ 0, & \text{otherwise} \end{cases}.$$

The second optimization objective f_2 represents risks. In order to penalize any changes to the timetable when the risks are high, the objective function is defined as

$$f_2 = \sum_{i=1}^{n} r_i \sum_{c \in C} \left|E_{c,i}\right|,$$

where r_i represents the risk at time interval i.

To penalize unnecessary energy fluctuations, the third objective f_3 averages the consecutive changes in energy amounts for all consumer loads, i.e.,

$$f_3 = \frac{1}{(n-1)|C|} \sum_{i=2}^{n} \sum_{c \in C} \left|E_{c,i} - E_{c,i-1}\right|.$$

2.4 Weighted Sum Approach

Since the optimal solutions to this problem appear to reside in the convex region of the objective space, we use a weighted sum approach to transform all objective values into a single one. The single objective function to be minimized thus equals

$$f = w_1 f_1 + w_2 f_2 + w_3 f_3$$

under the condition $w_1 + w_2 = 1$. The weight w_3 can be set independently of w_1 and w_2 and serves as a measure of limiting the energy fluctuations.

3 Optimization Approach

3.1 Setting Weights in the Weighted Sum

To obtain diverse solutions with the weighted sum approach, a good strategy for setting the weights is needed. While we plan to use a more sophisticated approach for this purpose in future work, these initial experiments were made by choosing equidistant values of w_1 from the interval $[0, 1]$ and defining w_2 as $1 - w_1$. In order to limit energy fluctuations, we set w_3 to 10^{-3}. Smaller weights proved insufficient in limiting the fluctuations while larger weights interfered with the first two objectives, which are more important than the third.

3.2 Linearization

Since all of the objective functions specified in Section 2.3 are either non-linear or contain non-linear parts, specific techniques are required to linearize these objectives and ensure the problem fits the MILP form. In particular, it is necessary to linearize the absolute value of a real variable, the product of a binary variable and a real variable, the minimum of two variables, along with other non-linear function conditions. We use standard approaches to achieve linearization for all these cases [9].

3.3 Tool and Solver

We use the OR-Tools Python library[2] to implement and solve the single-objective MILP problem. The library is a comprehensive tool for solving optimization problems, including linear programming, integer programming, and combinatorial optimization. Specifically, we use the SCIP (Solving Constraint Integer Programs) solver [1] integrated within OR-Tools[3] for solving MILP problem instances.

To solve a MILP problem using OR-Tools and the integrated SCIP solver, the following steps are performed: import the linear solver wrapper, declare the SCIP solver, define the variables with their respective bounds, set the constraints and the objective function and lastly, analyze and display the solution.

4 Experiments

We first conduct experiments using a basic scenario with a single consumer load. Then, we variate some parameters of this scenario to see how they affect the resulting solutions.

4.1 Experimental Setup

The basic scenario has the following parameters:

- Time is represented as 28 15-minute intervals.
- The demand period starts at $i = 13$ and ends at $i = 16$.
- The total required reduction E^{T} equals -8 kWh and the required reduction E^{D} at each interval equals -2 kWh.
- The adjustment period has a duration of four intervals.
- The load change needs to be within $[-3\,\mathrm{kWh}, 3\,\mathrm{kWh}]$ for each interval $i = 5, 6, \ldots, 24$ and is fixed to 0 kWh for the remaining intervals.
- The forecast timetable energy E_i^{F} is constant and equals 12 kWh for all time intervals.
- The total energy constraint is unbounded.
- The risk equals 0.50 for all time intervals.
- All prices are constant: $p_i = 0.25$ EUR, $p^{\mathrm{R}} = 0.50$ EUR and $p^{\mathrm{P}} = 1.00$ EUR.

[2] https://developers.google.com/optimization

[3] https://github.com/google/or-tools/blob/stable/ortools/linear_solver/samples/mip_var_array.py

The three scenario variants differ from the basic as follows. The first scenario variant has no demand. In the second and third scenario variant, the total energy change is set to 0 kWh ensuring the reduction in energy consumption in some intervals is matched with its increase in others. Additionally, the third scenario variant has no adjustment period, i.e, $n_{\mathrm{A}} = 0$.

4.2 Results and Discussion

We discuss here the results of our original scenario and its three variants. They are depicted also in plots in Figures 1 to 4, which show with a black line how the consumer load changes from its planned timetable. Consumer load flexibility at each time interval is shown in gray (there is no flexibility in the first four and last four intervals). The demand period is denoted in red and the adjustment period in blue. In most cases (unless the risk has a large weight), the consumer reduces the load in the demand period enough to meet the required demand and earn the entire available rebate while not incurring any penalty. The amount of this reduction and the energy change outside of this period differ for the various scenario variants.

4.2.1 Original Scenario. When the risk has a large weight, the load does not change outside of the demand period (see the top plot in Figure 1). However, when the impact of risk is minimal (bottom plot in Figure 1), the load is reduced everywhere except during the adjustment period. This strategy artificially increases the perceived load reduction to maximize the rebate, as dictated by the rebate calculation formula.

4.2.2 Scenario Variant #1: No Demand. If the optimization is called without a demand, the result depends on the weighting of the first two objectives. As long as the impact of risk is significant (top plot in Figure 2), the load does not change. Otherwise, the load is reduced to the maximum extent in each interval (bottom plot in Figure 2). This approach minimizes the function f_E, therefore reducing costs. This means that the consumer behavior can change when optimized even if no demand is present.

4.2.3 Scenario Variant #2: Zero Total Energy Change. Due to the zero energy constraint, the consumer makes adjustments solely within the demand and adjustment periods (see Figure 3). During the adjustment period, the user offsets the consumption from the demand period, thereby achieving a maximal rebate. To adhere to the requirement of minimizing risks and fluctuations in other intervals, no additional changes are made, as such actions would increase the objective value.

4.2.4 Scenario Variant #3: Zero Total Energy Change and No Adjustment Period. When the baseline is not adjusted, the load is increased in intervals outside of the demand period, regardless whether they occur before or after it. The specific intervals when this happens depend on the solver and are random as they lead to the same objective function value. An example of such a case in depicted in Figure 4.

The last two variants additionally confirm that the usage of the adjustment period enables exploitation – the entire rebate can be gained with a smaller load reduction in the demand period if the load is increased in the adjustment period.

5 Conclusions

This paper focuses on demand response optimization and the growing role of prosumers in energy systems. A standard MILP framework is used to set the consumer load energies within

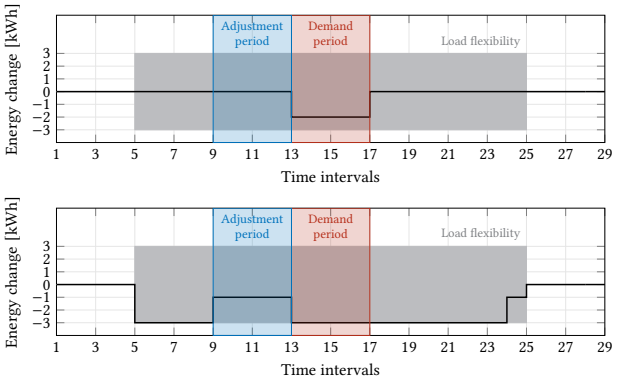


Figure 1: Results for the original scenario with $w_1 = 0.6$ and $w_2 = 0.4$ (top) and $w_1 = 0.8$ and $w_2 = 0.2$ (bottom).

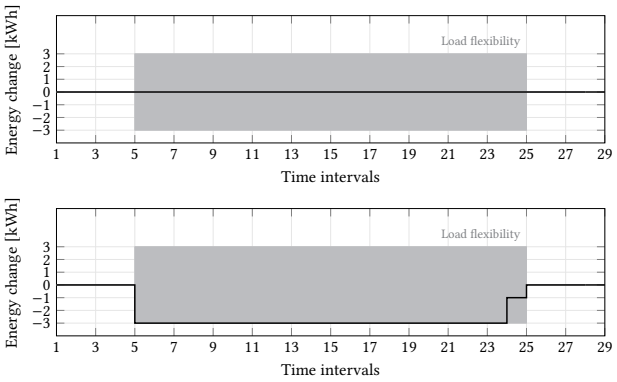


Figure 2: Results for the variant without demand with $w_1 = 0.5$ and $w_2 = 0.5$ (top) and $w_1 = 0.7$ and $w_2 = 0.3$ (bottom).

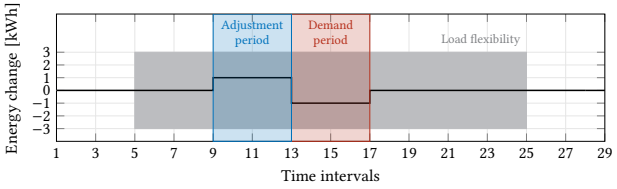


Figure 3: Results for the variant with zero total energy change with $w_1 = 0.6$ and $w_2 = 0.4$.

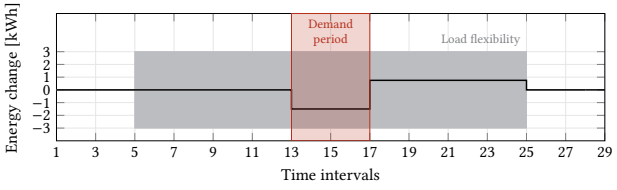


Figure 4: Results for the variant with zero total energy change and no adjustment period with $w_1 = 0.6$ and $w_2 = 0.4$.

their flexibility so that the costs, risks and energy fluctuations are all minimized. Since the objectives are scalarized with the weighted sum approach, correctly setting their weights is crucial for generating a set of diverse solutions representing various trade-offs between costs and risks.

By creating three scenario variants, we were able to explore the effect of some parameters on the optimization outcome. We observe that:

- Regardless of the variant, the optimal load schedule does not deviate from the forecast one if the importance of risk is too high, i.e., if the weight w_2 is too large. This critical value of w_2 depends on the scenario variant.
- If the consumer is obliged to a zero sum in load increase and reduction, the optimal solution uses the minimal necessary resources to earn a rebate while avoiding excessive energy changes.
- When the adjustment period is unspecified, the prosumer is less likely to increase the load just before the demand period.

Moving forward, we need to refine the objectives. The current method to assess the baseline consumption is susceptible to exploitation and should be amended. We could calculate the consumer baseline from similar consumers that do not participate in DR as suggested in [6]. We will also need to revise the penalty calculation to account for the imminent change of tariffs in the Slovenian energy market. We additionally plan to improve the calculation of risks to ensure more robust optimization and real-world applicability. Finally, we intend to develop a better strategy for setting the weights, targeting values with the most significant impact rather than evenly distributing them.

Acknowledgements

The SEEDS project is co-funded by the European Union's Horizon Europe innovation actions programme under the Grant Agreement n°101138211. The authors acknowledge the financial support from the Slovenian Research and Innovation Agency (research core funding No. P2-0209). The authors wish to thank Bernard Ženko, Martin Žnidaršič and Aljaž Osojnik for helpful discussions when shaping this work.

References

[1] Tobias Achterberg. 2009. SCIP: Solving constraint integer programs. *Mathematical Programming Computation*, 1, 1–41. DOI: 10.1007/s12532-008-0001-1.

[2] AEIC Load Research Committee. 2009. Demand response measurement & verification: Applications for load research. Tech. rep. AEIC Load Research Committee.

[3] István Balázs, Attila Fodor, and Attila Magyar. 2021. Quantification of the flexibility of residential prosumers. *Energies*, 14, 4860. DOI: 10.3390/en14164860.

[4] Martina Capone, Elisa Guelpa, and Verda Vittorio. 2021. Multi-objective optimization of district energy systems with demand response. *Energy*, 227, 120472. DOI: 10.1016/j.energy.2021.120472.

[5] Antonio Gabaldón, Ana García-Garre, María Carmen Ruiz-Abellón, Antonio Guillamón, Carlos Álvarez-Bel, and Luis Alfredo Fernandez-Jimenez. 2021. Improvement of customer baselines for the evaluation of demand response through the use of physically-based load models. *Utilities Policy*, 70, 101213. DOI: 10.1016/j.jup.2021.101213.

[6] Joe Glass, Stephen Suffian, Adam Scheer, and Carmen Best. 2022. Demand response advanced measurement methodology: Analysis of open-source baseline and comparison group methods to enable CAISO demand response resource performance evaluation. Tech. rep. California Independent System Operator (CAISO).

[7] Pedro L. Magalhães and Carlos Henggeler Antunes. 2020. Comparison of thermal load models for MILP-based demand response planning. In *Sustainable Energy for Smart Cities*. Springer International Publishing, Cham, 110–124.

[8] Javier Parra-Domínguez, Esteban Sánchez, and Ángel Ordóñez. 2023. The prosumer: A systematic review of the new paradigm in energy and sustainable development. *Sustainability*, 15, 13. DOI: 10.3390/su151310552.

[9] Nace Sever. 2022. *Časovno razporejanje terenskih nalog z mešanim celoštevilskim linearnim programiranjem*. Bachelor's Thesis. University of Ljubljana, Faculty of Mathematics and Physics. https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=slv&id=140427.

Predicting Hydrogen Adsorption Energies on Platinum Nanoparticles and Surfaces with Machine Learning

Lea Gašparič
lea.gasparic@ijs.si
Jožef Stefan Institute, Jožef Stefan international postgraduate schoole
Ljubljana, Slovenia

Anton Kokalj
tone.kokalj@ijs.si
Jožef Stefan Institute, Jožef Stefan international postgraduate school
Ljubljana, Slovenia

Sašo Džeroski
saso.dzeroski@ijs.si
Jožef Stefan Institute, Jožef Stefan international postgraduate school
Ljubljana, Slovenia

Abstract

The growing interest in hydrogen gas as a fuel drives research into environmentally friendly hydrogen production methods. One viable approach of obtaining hydrogen is the electrocatalysis of water, which includes the hydrogen evolution reaction (HER) as one of the half-reactions. In the search of highly active catalysts for the HER, machine learning can be effectively utilized to develop models for calculating hydrogen adsorption energy, a key descriptor of catalytic activity. In this study, we learned models for predicting hydrogen adsorption energy on platinum. We used various machine-learning (ML) techniques on two datasets, one for extended surfaces and the other for nanoparticles. The respective results reveal that ML models for extended surfaces are more accurate than those for nanoparticles, and that the features describing the local environment are the most significant for the predictions. For surfaces, the coordination number is the most relevant feature, while the d-band center is the most important for nanoparticles. The ML models developed in this study lack sufficient accuracy to provide reliable results, highlighting the need for further investigation with additional features or larger datasets.

Keywords

platinum, hydrogen, DFT calculations, decision trees, feature ranking

1 Introduction

A lot of scientific and societal interest is devoted to hydrogen fuel, which can generate electrical power by producing water as a byproduct. One environmentally friendly method of producing hydrogen is through the electrocatalysis of water, where hydrogen and oxygen gases are formed. This process involves two reactions: oxygen and hydrogen evolution reactions. Considerable effort is being directed towards improving catalysts for both reactions and understanding the fundamental processes involved [21, 13]. In this contribution, we will focus on the hydrogen evolution reaction (HER), for which platinum is known to be a highly active catalyst due to its near-optimal hydrogen adsorption free energy [15, 21]. However, the high cost of platinum motivates ongoing research of alternative materials.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.8689

The mechanism of HER includes adsorbed hydrogen atom (H*) as an intermediate. Consequently, the adsorption energy of hydrogen is often used as a descriptor of the catalytic activity of the material [15, 21]. The most straightforward approach to obtain the adsorption energies is with density-functional theory (DFT) calculations. However, as the size of the system and the number of different adsorption sites increase, a full DFT analysis becomes computationally unfeasible. To address this challenge, machine-learning methods can be employed to predict hydrogen adsorption energies based on DFT results, enabling the investigation of more complex systems [10]. For example, bimetallic nanoparticles were investigated by Jäger et al. [8] and Zhang et al. investigated amorphous systems [20].

This contribution focuses on the use of machine learning for predicting hydrogen adsorption energies on platinum using electronic and geometric descriptors. Two separate datasets were constructed, one for surfaces and the other for nanoparticles. By employing supervised learning and attribute ranking, we built ML models, assessed their accuracy and analyzed whether the two datasets exhibit similar correlations. The idea of the contribution is illustrated in Figure 1.

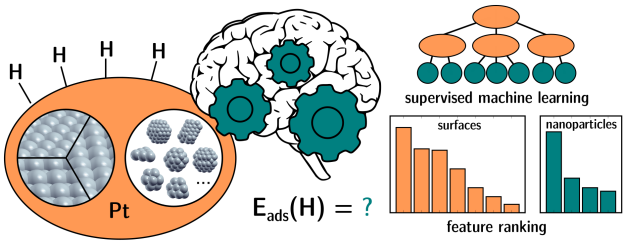


Figure 1: Supervised machine learning and feature ranking was performed for hydrogen adsorption energy on platinum catalysts modeled as surfaces and nanoparticles.

2 Materials and Methods

2.1 DFT Calculations and Datasets

We utilized DFT calculations to calculate hydrogen adsorption energies (a target variable for ML) and electronic descriptors for ML. We also utilized geometric descriptors. Two datasets were constructed, one for platinum nanoparticles and the other for platinum surfaces.

DFT calculations were performed with the Perdew-Burke-Ernzerhof (PBE) approximation [17], a plane-wave basis set, and PAW pseudopotentials [3]. Energy cutoffs were set to 50 and 575 Ry for wavefunctions and electron density,

respectively. Methfessel-Paxton smearing [12] of 0.02 eV was employed.

Pt(111), Pt(100), and Pt(110) surface slab models were constructed with the calculated lattice parameter of bulk Pt (3.97 Å). The models of Pt(111) and Pt(100) surfaces consist of 4 atomic layers, with the bottom layer fixed to bulk positions, while Pt(110) has 6 atomic layers with the bottom two layers fixed. To achieve a greater variety of adsorption sites, Pt(111) and Pt(100) were also modeled with a missing-row defect. All surface models are shown in Figure 2. Calculations accounted for the dipole correction and periodic images of slabs were separated by at least 15 Å of vacuum. Different sizes of surface supercells were used, and the k-point grid for (1×1) surface unit cells of Pt(111), Pt(100), and Pt(110) were 12×12×1, 11×11×1, and 11×8×1, respectively. For larger supercells, the number of k-points was adapted accordingly.

Calculations with nanoparticles were performed with the gamma k-point and Martyna-Tuckerman correction for isolated systems [11]. Nanoparticles were modeled with different shapes and sizes, consisting of 3 and up to 116 atoms. Their periodic images were separated by at least 15 Å of vacuum. All calculations were preformed with the Quantum ESPRESSO package [5].

The hydrogen adsorption energy was calculated as:

$$E_{\mathrm{ads}} = E_{\mathrm{H}*} - E_{*} - \frac{1}{2}E_{\mathrm{H_2}} \tag{1}$$

where $E_{\mathrm{H}*}$ is the calculated energy of optimized adsorption system, E_* is the energy of the standalone platinum system, and $E_{\mathrm{H_2}}$ is the energy of the hydrogen molecule. All performed calculations included only one adsorbed H atom per supercell or nanoparticle.

As an electronic descriptor, we used the d-band center, which is considered to be a good indicator of metal reactivity [6]. It was obtained through DFT calculations using the following equation:

$$\varepsilon_{\mathrm{d}} = \frac{\int_{-\infty}^{\infty} n_{\mathrm{d}}(E)EdE}{\int_{-\infty}^{\infty} n_{\mathrm{d}}(E)dE} \tag{2}$$

where E is the energy and n_{d} is the projected density of states on d-orbitals of the atoms forming the adsorption site.

For the geometric descriptors, we determined the average coordination number of Pt atoms forming the adsorption site, as well as the generalized coordination number (GCN) of the adsorption site [2], calculated as:

$$\mathrm{GCN}(i) = \sum_{j=1}^{N_i} \frac{\mathrm{CN}(j)}{\mathrm{CN}_{\mathrm{max}}} \tag{3}$$

where i denotes an atom or a group of atoms forming the adsorption site, N_i is the number of first nearest neighbors of i, which are denoted with j. $\mathrm{CN}(j)$ is the coordination number of atom j and $\mathrm{CN}_{\mathrm{max}}$ is the maximal coordination of a given site found in the bulk material.

In addition, the type of adsorption site was used as a descriptor. For extended surfaces, the coverage of H atoms, the surface area per H atom and surface type were also used for learning. For nanoparticles, some descriptors relevant to the size of nanoparticles were also utilized, in particular: the number of all atoms (N_{all}) in the nanoparticle, the number of surface atoms (N_{surf}), the maximal (r_{max}) and minimal (r_{min}) distances from the center of the nanoparticle to the surface atoms and the distance from the center of the nanoparticle to the adsorption site (r_{ads}). The datasets for surfaces and nanoparticles contained 46 and 85 data points, respectively.

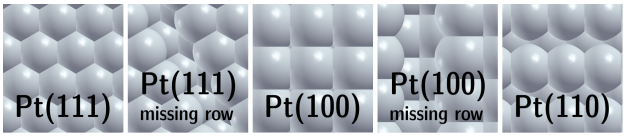


Figure 2: Models of extended surfaces used to calculate hydrogen adsorption energies.

2.2 Machine-Learning Methods

The prepared datasets were analyzed using the Weka software package [4]. The target value in both datasets is the hydrogen adsorption energy, making this a regression task. Supervised machine learning was employed to develop models for predicting the target value, which were evaluated by 10-fold cross-validation.

One of the used methods is linear regression, that computes the linear relationship between the target value and the descriptors. The relevant descriptors included in the equation were selected according to the M5 method [18]. This method iteratively removes descriptors with the smallest effect on the model until the error of the model no longer decreases.

We also used the random forest method [7, 1] with 100 trees of unlimited depth. With this method, multiple decision trees were constructed by selecting relevant features from a random subset of $\mathrm{int}(\log_2(m)+1)$ features, where m is the total number of features. The final values are the averages of the predictions from the individual trees.

To obtain an explainable ML model, we also built regression trees using the M5' method [18, 19]. In this method, trees are built by splitting the training sets according to attributes that maximize the standard deviation reduction. After the trees are constructed, they are pruned to avoid overfitting and smoothed to address discontinuities between the leaves. For our datasets, we used unpruned trees to prevent the formation of trees that are too small and give poor predictions. We also restricted tree branching to a minimum of 6 instances per leaf node for surfaces and 20 for nanoparticles to avoid overfitting the data and to ensure trees of sufficient size.

We also performed variable importance estimation and ranking for our selected descriptors with all data points used as a test set. To evaluate the importance of the descriptors with respect to hydrogen adsorption energy, we employed two methods: ReliefF [9] and correlation [16]. The ReliefF method is more sensitive to feature interactions and works by calculating the distances between training instances and identifying the 'nearest hit' and 'nearest miss'. It then adjusts the weights of the differing descriptors between the target and nearest instances. The correlation method evaluates the Pearson correlation coefficient [16]

between the features and the target variable, without accounting for interactions between features. It gives scores ranging from −1 to 1, with 1 being the highest correlation score, a score of −1 indicates anti-correlation, and 0 indicates no correlation.

3 Results and Discussion

3.1 Machine-Learning Models

Supervised machine learning was performed using linear regression, random forest, and M5' regression tree. The obtained Pearson's correlation coefficients and root mean squared errors (RMSE) between true and predicted values are shown in Table 1.

We can observe that not all ML models provide better RMSE values compared to those calculated with a simple arithmetic average, referred to as the default predictor. For surfaces, linear regression and random forest perform the best and yield similar results. The regression tree model performs the worst and has higher RMSE compared to the default predictor. For nanoparticles, all methods yield errors close to those of the default predictor and correlation coefficients bellow 0.5.

The obtained results indicate that with the selected descriptors, the hydrogen adsorption energies are more accurately predicted on surfaces, which are simpler as compared to nanoparticles. Surfaces have high symmetry and only a handful of different adsorption sites, while nanoparticles have different shapes and sizes, consist of different facets, and each nanoparticle has numerous different adsorption sites. This gives a huge variety of adsorption sites that can make the prediction of adsorption energies harder.

Considering the best models, the obtained adsorption energies have an error of ±0.13 eV for surfaces and ±0.22 eV for nanoparticles. Due to the exponential dependence of reaction rate and adsorption energy, even a small error in adsorption energy hugely affects the reaction rate. Hence, the models, particularly for nanoparticles, do not provide sufficiently accurate results for any practical use.

The selected ML models also provide insights into the relations between the considered features and the target variable. The linear regression model for nanoparticles includes only the d-band center and a factor for the hollow adsorption site, whereas the equation for surfaces is more complex. It includes adsorption site, surface type, and both coordination numbers. This indicates that for nanoparticles, the d-band center is the most relevant factor, while for surfaces, geometric factors exhibit greater predictive value.

Table 1: Pearson's correlation coefficients (CC) and root mean squared errors (RMSE) in eV units for all three used ML methods. For comparison, RMSE of the default predictor is also given.

	surfaces		Nanoparticles	
	CC	RMSE	CC	RMSE
linear regression	0.71	0.13	0.38	0.22
random forest	0.69	0.13	0.34	0.22
M5' decision tree	0.49	0.19	0.34	0.22
default predictor	/	0.18	/	0.23

The regression-tree models shown in Figure 3 have lower accuracy and, consequently, are less reliable.

The ML models could be improved by expanding the dataset or by calculating additional descriptors. For surfaces, more data can be obtained through calculations on a wider variety of surface types and by accounting for different surface defects. However, expanding the dataset for nanoparticles is limited by their size, since DFT calculations for larger particles are computationally too demanding. Therefore, a larger number of different smaller particles can be tested instead. Using more sophisticated descriptors such as atom-centered symmetry functions, smooth overlap of atomic positions and many body tensor representation could also improve the results, but would require different sampling of adsorption structures. The use of transfer learning from pre-trained models based on chemical structures could also lead to significant improvements.

3.2 Feature Ranking

Feature ranking was performed for both surfaces and nanoparticles, with the results presented in Figure 4. The ReliefF and correlation importance criteria provide different rankings of features. For surfaces, the coordination number is identified as the most relevant descriptor, followed by the generalized coordination number. In contrast, for nanoparticles, the d-band center is the most important descriptor. Features describing the size of different nanoparticles show lower relevance for predictions. The most relevant features in both data sets describe the local environment of the adsorption site, indicating the local nature of adsorption.

The importance of the d-band center is already well-documented in the literature [14], as it correlates with the reactivity of metals. As seen from the graphs, the d-band center is not so strongly correlated with the hydrogen binding energy on surfaces. This can be attributed to the fact that on a perfectly flat surface, all surface atoms have the same d-band center. In contrast, on nanoparticles, the d-band center varies for each adsorption site because the atoms are not equivalent. Therefore, the d-band center is expected to be more relevant for nanoparticles. For the ranking based on correlation, the calculated factors for the

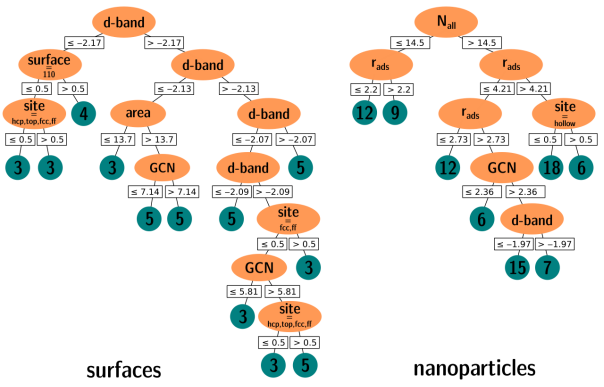


Figure 3: Schematic representation the obtained random-tree models for ideal surfaces and nanoparticles. Nodes are denoted with orange and the resulting classes are represented with turquoise circles and include the number of data points in the class.

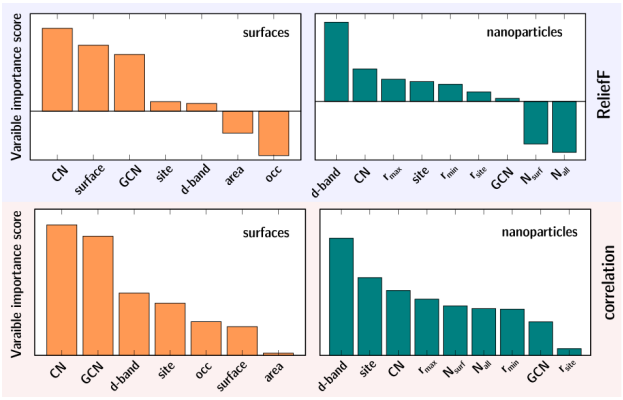


Figure 4: Variable importance scores calculated by the ReliefF and correlation criteria. Importance scores for correlation criteria are given as absolute values.

d-band center are negative. This indicates that a lower d-band center corresponds to a higher adsorption energy and consequently a less reactive site, which is physically intuitive.

It is also interesting to note that the surface type descriptor is not very relevant according to correlation, yet it becomes the second most important feature when other descriptor are considered. This can be attributed to the fact that this descriptor has the same value for all adsorption sites on the same surface. However, when combined with other descriptors, it can give additional information, as similar adsorption sites on different surfaces can yield considerably different adsorption energies.

4 Conclusion

We applied different ML techniques to predict the adsorption energy of hydrogen on platinum surfaces and nanoparticles using simple geometric and electronic descriptors. Models for predicting adsorption energy on surfaces performed better, with the linear regression and random forest methods showing the highest correlation coefficient and accuracy. In contrast, predictions for nanoparticles yielded lower correlation coefficients and accuracy similar to the one calculated by a default predictor. Therefore, the models presented in this contribution do not provide accurate estimation of hydrogen adsorption energies. Utilizing more sophisticated descriptors and larger training data sets could enhance the performance of these models.

Differences between datasets are also evident in feature ranking. For surfaces, coordination numbers are the most relevant descriptors, while for nanoparticles, the d-band center shows the highest relevance. All these relevant descriptors are related to the local environment of the adsorption site, indicating that adsorption is a local phenomenon.

References

[1] Leo Breiman. 2001. Random forests. *Machine Learning*, 45, 1, (Oct. 2001), 5–32. DOI: 10.1023/A:1010933404324.

[2] Federico Calle-Vallejo, José I. Martínez, Juan M. García-Lastra, Philippe Sautet, and David Loffreda. 2014. Fast prediction of adsorption properties for platinum nanocatalysts with generalized coordination numbers. *Angew. Chem. Int. Ed.*, 53, 32, (Aug. 2014), 8316–8319. DOI: 10.1002/anie.201402958.

[3] Andrea Dal Corso. 2014. Pseudopotentials periodic table: From H to Pu. *Comput. Mater. Sci.*, 95, (Dec. 2014), 337–350. (files: H.pbe-kjpaw_psl.1.0.0.UPF, Pt.pbe-n-kjpaw_psl.1.0.0.UPF). DOI: 10.1016/j.commatsci.2014.07.043.

[4] Eibe Frank, Mark A. Hall, and Ian H. Witten. 2016. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques". Fourth Edition.* Morgan Kaufmann. https://ml.cms.waikato.ac.nz/weka/Witten_et_al_2016_appendix.pdf.

[5] Paolo Giannozzi et al. 2009. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys: Condens. Matter*, 21, 39, 395502. Code available from http://www.quantum-espresso.org/. DOI: 10.1088/0953-8984/21/39/395502.

[6] Bjørk Hammer and Jens K. Nørskov. 1995. Electronic factors determining the reactivity of metal surfaces. *Surf. Sci.*, 343, 3, (Dec. 1995), 211–220. DOI: 10.1016/0039-6028(96)80007-0.

[7] Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE, 278–282.

[8] Marc O. J. Jäger, Yashasvi S. Ranawat, Filippo Federici Canova, Eiaki V. Morooka, and Adam S. Foster. 2020. Efficient machine-learning-aided screening of hydrogen adsorption on bimetallic nanoclusters. *ACS Comb. Sci.*, 22, 12, (Dec. 2020), 768–781. DOI: 10.1021/acscombsci.0c00102.

[9] Igor Kononenko, Edvard Šimec, and Marko Robnik-Šikonja. 1997. Overcoming the myopia of inductive learning algorithms with RELIEFF. *Applied Intelligence*, 7, 1, (Jan. 1997), 39–55. DOI: 10.1023/A:1008280620621.

[10] Jin Li et al. 2023. Machine learning-assisted low-dimensional electrocatalysts design for hydrogen evolution reaction. *Nano-Micro Lett.*, 15, 1, (Oct. 2023), 227–27. DOI: 10.1007/s40820-023-01192-5.

[11] Glenn J. Martyna and Mark E. Tuckerman. 1999. A reciprocal space based method for treating long range interactions in ab initio and force-field-based calculations in clusters. *J. Chem. Phys.*, 110, 6, (Feb. 1999), 2810–2821. DOI: 10.1063/1.477923.

[12] Michael Methfessel and Anthony Thomas Paxton. 1989. High-precision sampling for brillouin-zone integration in metals. *Phys. Rev. B*, 40, 6, (Aug. 1989), 3616–3621. DOI: 10.1103/PhysRevB.40.3616.

[13] Bishnupad Mohanty, Piyali Bhanja, and Bikash Kumar Jena. 2022. An overview on advances in design and development of materials for electrochemical generation of hydrogen and oxygen. *Mater. Today Energy*, 23, (Jan. 2022), 100902. DOI: 10.1016/j.mtener.2021.100902.

[14] Anders Nilsson, Lars G. M. Pettersson, Bjørk Hammer, Thomas Bligaard, Claus Hviid Christensen, and Jens K. Nørskov. 2005. The electronic structure effect in heterogeneous catalysis. *Catal. Lett.*, 100, 3, (Apr. 2005), 111–114. DOI: 10.1007/s10562-004-3434-9.

[15] Jens Kehlet Nørskov, Thomas Bligaard, Ashildur Logadottir, John R. Kitchin, Jingguang G. Chen, Stanislav Pandelov, and Ulrich Stimming. 2005. Trends in the exchange current for hydrogen evolution. *J. Electrochem. Soc.*, 152, 3, (Jan. 2005), J23. DOI: 10.1149/1.1856988.

[16] Karl Pearson. 1895. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58, 347-352, 240–242.

[17] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. 1996. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77, 18, (Oct. 1996), 3865–3868. DOI: 10.1103/PhysRevLett.77.3865.

[18] John R et al. Quinlan. 1992. Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*. Vol. 92. World Scientific, 343–348.

[19] Yong Wang and Ian H Witten. 1997. Inducing model trees for continuous classes. In *Proceedings of the ninth European conference on machine learning* number 1. Vol. 9. Citeseer, 128–137.

[20] Jiawei Zhang, Peijun Hu, and Haifeng Wang. 2020. Amorphous catalysis: machine learning driven high-throughput screening of superior active site for hydrogen evolution reaction. *J. Phys. Chem. C*, 124, 19, (May 2020), 10483–10494. DOI: 10.1021/acs.jpcc.0c00406.

[21] Jing Zhu, Liangsheng Hu, Pengxiang Zhao, Lawrence Yoon Suk Lee, and Kwok-Yin Wong. 2020. Recent advances in electrocatalytic hydrogen evolution using nanoparticles. *Chem. Rev.*, 120, 2, (Jan. 2020), 851–918. DOI: 10.1021/acs.chemrev.9b00248.

SmartCHANGE Risk Prediction Tool: Demonstrating Risk Assessment for Children and Youth

Marko Jordan
Jožef Stefan Institute,
Department of Intelligent Systems
Ljubljana, Slovenia
marko.jordan@ijs.si

Nina Reščič
Jožef Stefan Institute,
Department of Intelligent Systems
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia
nina.rescic@ijs.si

Sebastjan Kramar
Jožef Stefan Institute,
Department of Intelligent Systems
Ljubljana, Slovenia
sebastjan.kramar@ijs.si

Marcel Založnik
Jožef Stefan Institute,
Department of Intelligent Systems
Ljubljana, Slovenia
marcel.zaloznik@ijs.si

Mitja Luštrek
Jožef Stefan Institute,
Department of Intelligent Systems
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia
mitja.lustrek@ijs.si

Abstract

Non-communicable diseases (NCDs) have become a significant public health challenge in developed countries, driven by common risk factors such as obesity, low physical activity, and unhealthy lifestyle choices. Early childhood and adolescence are crucial for establishing healthy behaviours, and early intervention can play a crucial role in preventing or delaying the onset of NCDs later in life. However, current tools for identifying high-risk individuals are primarily designed for adults, which results in missed early detection opportunities in younger populations. The SmartCHANGE project (https://smart-change.eu/) seeks to bridge this gap by developing reliable AI tools that assess risk factors in children and adolescents as accurately as possible while promoting optimized risk reduction strategies.

In developing the risk assessment tool, we addressed the challenge of merging diverse datasets, predicting missing data to create longitudinal datasets, implementing existing validated models for diabetes (QxMD) and cardiovascular disease (SCORE2), and ultimately creating a simple online application to demonstrate the functionality of the developed risk tool.

Keywords

risk tool, dataset merge, neural networks, online application

1 Introduction

In developed countries, non-communicable chronic diseases (NCDs) have emerged as the foremost public health challenge over recent decades. According to the World Health Organization (WHO), NCDs account for more than 70% of mortality in the European region [18]. Common risk factors for NCD include obesity, poor physical fitness, and unhealthy lifestyle habits such as inadequate physical activity, sedentary behaviour, poor nutrition, insufficient sleep, smoking, and excessive alcohol consumption. Embracing a healthy lifestyle can improve physical, social, and mental well-being, especially among youth, while mitigating the risks of NCD-related morbidity and mortality [15], [14], [5].

Traditionally, clinical prevention strategies for NCDs have been directed at adults, as the risk factors typically become evident in adulthood. However, recent evidence suggests that focusing interventions on children and adolescents can be a more effective strategy for reducing NCD risk through behaviour modification [13]. While NCDs may not appear in childhood or adolescence, early signs can alreadexistnt. Tackling risk factors and promoting healthy habits during these stages can prevent or delay NCDs later in life [12]. Childhood and youth are also crucial periods for establishing healthy lifestyle habits. Since risk factors for NCDs often persist from childhood into adulthood [9], early risk assessment and reduction of risk factors can potentially lower the incidence of NCD. Lastly, NCDs in youth are a significant global health challenge, with nearly one in five adolescents worldwide being overweight or obese [1].

Identifying high-risk individuals for future health problems is essential for targeted preventive interventions. Existing tools focus mainly on adults [6], for instance predicting 10-year risk of developing cardiovascular disseased [17] or diabetes [8], missing the opportunity to identify high-risk individuals during childhood and adolescence, a critical period for forming lifestyle habits. However, recognition of health risks is not a trivial task. For instance, only 35% of doctors in the UK are aware of the recommendations for physical activity, and only 13% can specify the recommended weekly duration. Moreover, more than 80% of parents of inactive children incorrectly believe that their children are sufficiently active [4]. Developing risk prediction tools for children and youth would significantly improve NCD prevention and promote cost-effective strategies.

This paper presents the development of an initial demo application of a risk assessment tool designed for children and adolescents in the SmartCHANGE project [3] - merging datasets, predicting missing data to build longitudinal datasets, and implementing existing validated models for diabetes (QxMD) and cardiovascular disease (SCORE2) and finally, the application development.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.8844

Table 1: Overview of Selected Datasets

Dataset Name	SLOfit	LGS	YFS	AFINA-TE
Country of Origin	SI	BE	FI	PT
Age Range	5 - 20	5 - 25	0 - 60	5 - 25
Longitudinal Study	Yes	Yes	Yes	No
# of Participants	280,165	17,991	3,596	1,632
# of Measurements	3,121,399	31,127	32,364	1,632
# of Variables	13	80	24	59
% of Missing Values	2.55%	16.25%	39.49%	33.53%

2 Methodology

2.1 Datasets

To estimate the risk of non-communicable diseases in children, ideally, one would need a dataset that tracks risk factors from a young age (when the prediction is made) to an older age (when these diseases typically emerge). Such comprehensive longitudinal datasets would allow for accurate predictions of an individual's likelihood of developing a disease later in life based on their early risk factors. However, such datasets are currently unavailable, so we must rely on a collection of partial and often heterogeneous datasets.

In our study, we have chosen 16 types of variables that are used by risk models SCORE2 [17] and QxMD [8]. The datasets we were using are described in Table 1. The SLOfit program is a school fitness monitoring initiative in Slovenia [11]. The Leuven Growth Study (LGS)[2, 16] is a longitudinal study initiated in 1969 that evaluates physical fitness. The Cardiovascular Risk in Young Finns Study (YFS)[10], started in the late 1970s, focuses on early cardiovascular disease risk factors. The AFINA-TE dataset [7] is part of an intervention program in Portugal designed to enhance physical fitness, activity, and nutritional knowledge among children and adolescents.

2.2 Data Imputation Through Datasets

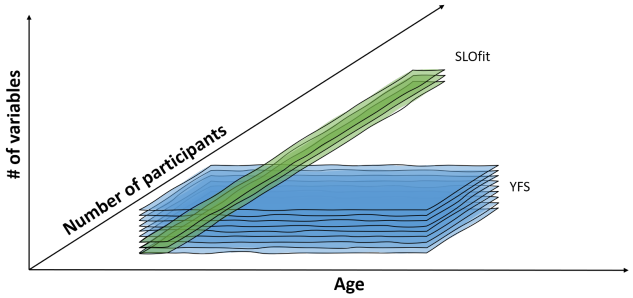
The first step involved imputing missing values within each dataset (see Figure 1 for representation). To guide this process, we calculated the coverage for each variable. Initially, we used only fully observed variables—such as height, weight, and sex—as features in models to impute missing values for other variables. The variables were imputed based on their coverage using machine learning on existing features. After this initial imputation sweep, we had a complete, though potentially imperfect, dataset. In the second sweep, we treated all columns as complete, incorporating the newly imputed values from the first sweep. This allowed us to train models with a more comprehensive dataset, improving the accuracy of the imputation.

2.3 Longitudinal Data Imputation

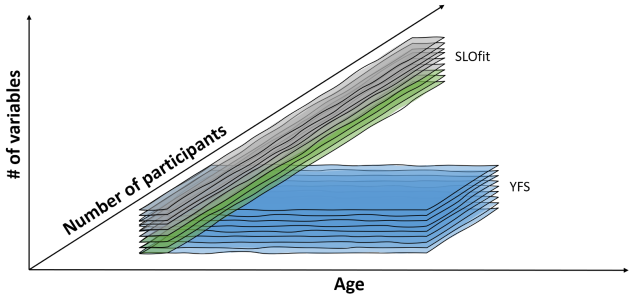
In the second step, we employed a similar approach but focused on merging the datasets to fill the new merged dataset longitudinally (see Figure 2 for representation). To maximize their overlap, we treated certain variables as equivalent—such as vertical jump from the LGS dataset and standing long jump from the SLOfit dataset.

Since the raw values of these variables differ, we standardized them by converting them to z-scores, which were calculated as follows:

$$z\_score = \frac{variable - mean}{standard\_deviation}.$$



(a) Example of the datasets pre-imputation.



(b) Example of the datasets post-imputation.

Figure 1: The YFS dataset (blue) covers a broad range of variables across a wide age span but includes a relatively small number of participants. In contrast, the SLOfit dataset (green) has many participants but includes fewer variables over a shorter age span. In the first step, we imputed the missing variables across the datasets (grey).

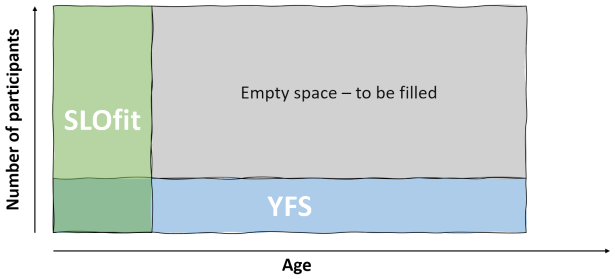


Figure 2: Longitudinal filling of the datasets.

For instance, a vertical jump one standard deviation above the mean in the LGS dataset was considered equivalent to a standing long jump one standard deviation above the mean in the SLOfit dataset. After matching and standardizing the columns across datasets, we merged the individual datasets into a single, comprehensive dataset and repeated the imputation process.

With a merged dataset free of missing values, we built models to predict attribute values at age 55—the oldest age supported by our data—using values from age 14. Due to the lack of data covering the entire age range from 14 to 55, we approached this in two stages: predicting from age 14 to 18 and then from 18 to

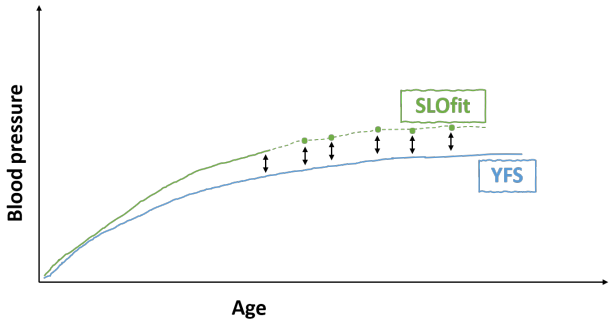


Figure 3: Population-based approach using z-scores.

55. The models used were simple neural networks with a single hidden layer.

This individual forecasting approach required available data for the same person from the start to the end age. However, since we had more data available for different people of various ages, we also explored a population-based approach to forecast the typical evolution of each variable. While this method is less personalized, it is also less prone to anomalies caused by atypical individuals. In the population-based approach, we again used z-scores, assuming that each person's z-score remains constant. For example, if someone's blood pressure is one standard deviation below the mean at age 14, it is assumed to stay one standard deviation below the mean at age 55 (see Figure 3).

2.4 Risk Models

The SCORE2 and QxMD models were used in the application to assess cardiovascular disease and type 2 diabetes risk. These models were chosen for their validity, robustness and effectiveness in predicting these chronic conditions. By incorporating both, healthcare practitioners can comprehensively evaluate cardiometabolic risk factors, aiding in well-informed patient management and intervention decisions.

The SCORE2 model, developed by the European Society of Cardiology, estimates the risk of cardiovascular events over ten years. It calculates the risk score by incorporating variables such as age, sex, smoking status, blood pressure, and lipid profile. Additionally, SCORE2 considers regional variations in risk factors, providing more accurate predictions tailored to specific populations [17].

The QxMD Diabetes Risk Calculator, a comprehensive clinical decision support tool, is employed to evaluate the risk of developing type 2 diabetes mellitus. This model integrates risk factors, including age, BMI, family history, physical activity level, and dietary habits, to estimate an individual's diabetes risk [8].

3 Evaluation

Table 2 presents the cross-validated evaluation results of our forecasting models. As anticipated, the errors in the first stage of individual forecasting are shallow due to the relatively short period. The emistakesin the second stage are higher but still considered acceptable, with the notable exceptions of weight and smoking. We hypothesize that the high variability during puberty, which many adolescents experience around age 14, complicates accurate weight forecasting. In population forecasting, the errors are generally more significant, which aligns with the less personalized nature of this method. However, weight is forecasted

	Ind. 18	Ind. 55	Pop.
Height [cm]	3.11	3.47	1.62
Weight [kg]	4.79	13.60	10.58
SBP [mmHg]	1.46	2.39	10.91
Total cholesterol [mmol/L]	0.05	0.10	0.64
HDL [mmol/L]	0.02	0.08	0.21
LDL [mmol/L]	0.05	0.17	0.51
Smoking [1-9]	1.01	1.72	2.26

Table 2: Mean absolute error for individual forecasting to ages 18 and 55, and for population forecasting.

with greater accuracy in this approach. In the future, we may explore combining both methods or select the more accurate one depending on the variable.

4 Demo Application

To show the general idea of the project, we constructed a demo application implemented with Python in the Dash framework. In the app, a user can specify the inputs (some inputs, such as steroid use, were fixed to make the app more concise) to the models, which in turn yielded two plots which showed how the cardiovascular and diabetes risk evolved from the currently selected age up to an age of an older adult, at age 55. In a different plot, we also showed how a risk factor chosen changes over time.

4.1 Risk Prediction using Demo Application

The developed demo application interface offers a dynamic tool for visualizing health risks based on various user-input parameters used in risk models (Figure 4). By allowing users to adjust these parameters, the dashboard generates real-time projections of two key risk metrics: a 10-year cardiovascular risk score and a 10-year risk of developing diabetes. These risks are shown in two line graphs, illustrating how these conditions' probability evolves with age. Additionally, the dashboard includes a feature that tracks the progression of a selected health parameter (BMI, systolic blood pressure, total cholesterol, HDL) over time, providing insight into how this factor might change as the individual ages. The developed tool intuitively explains how lifestyle and physiological factors contribute to long-term health risks, offering valuable insights for clinical decision-making and personal health management.

4.2 Further Development of the Application

The current version of the demo application is developed based on the data and models currently available. However, there remains an open question regarding the specific needs and preferences of the medical experts who will ultimately use the final application. To address this, we plan to present the current version to these experts and, based on their feedback, refine and enhance the application in subsequent iterations.

5 Conclusion

The SmartCHANGE project represents a significant step toward improving the early detection and prevention of non-communicable diseases (NCDs) in children and youth. While the tool presented in this paper is a demo version demonstrating some basic functionalities, our future work will focus on developing a more comprehensive web application for medical professionals and a mobile application for families. We also plan to enhance the tool

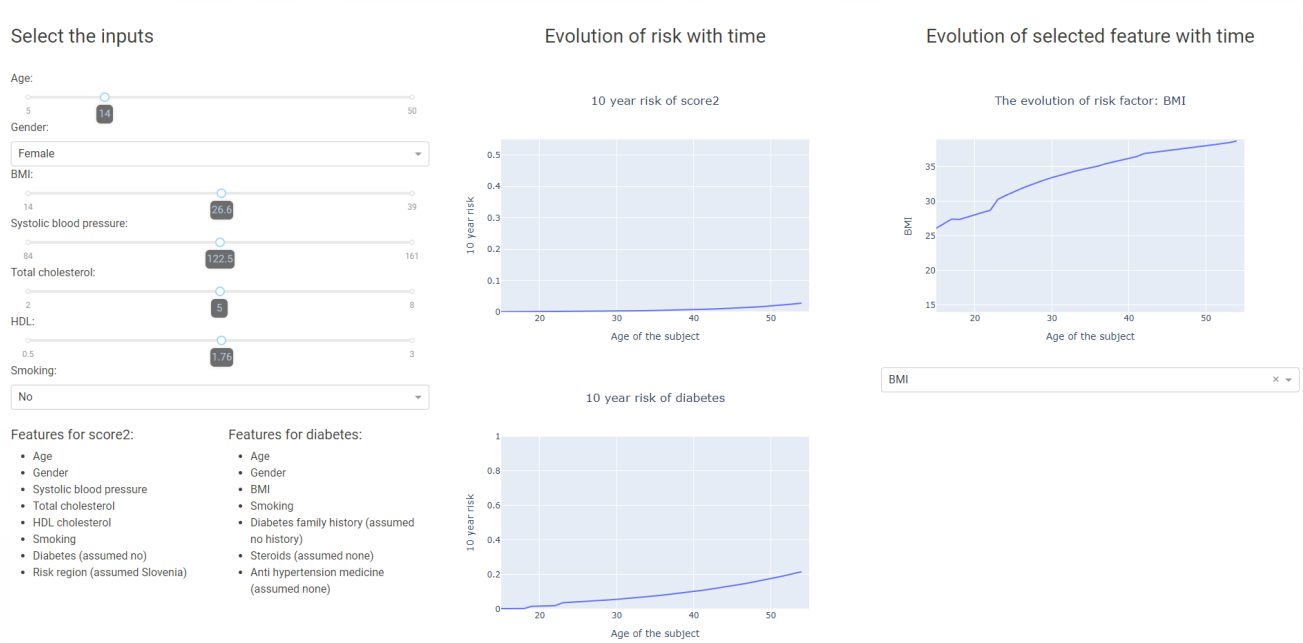


Figure 4: The figure is a dashboard interface that allows users to input various health-related parameters and observe the evolution of associated risks over time.

by replacing the current SCORE2 and QxMD risk models with more advanced models—Test2Prevent for diabetes and Healthy Heart Score for cardiovascular disease—incorporating features related to diet and physical activity. Additionally, the application will be updated to meet medical experts' needs based on their feedback.

Acknowledgements

This work was carried out as a part of the SmartCHANGE project, which received funding from the European Union's Horizon Europe research and innovation program under grant agreement No 101080965. The SLOfit dataset for wasvided by the University of Ljubljana (courtesy of Gregor Jurak et al.), the LGS dataset was provided by KU Leuven (courtesy of Martine ThomThomashe AFINA-TE dataset was provided by the University of Porto (courtesy of José Ribeiro) and the the University of Turku provided the YFS dataset are grateful for their support.

References

[1] P. S. Azzopardi, S. J. C. Hearps, K. L. Francis, E. C. Kennedy, A. H. Mokdad, N. J. Kassebaum, S. Lim, and et al. 2019. Progress in adolescent health and wellbeing: tracking 12 headline indicators for 195 countries and territories, 1990-2016. *Lancet*, 393, 10190, (Mar. 2019), 1101–1120.

[2] Gaston P Beunen, Robert M Malina, Marc A Van't Hof, Jan Simons, Michel Ostyn, Roland Renson, and Dirk Van Gerven. 1988. *Adolescent growth and motor performance: A longitudinal study of Belgian boys*. Human Kinetics Publishers.

[3] SmartCHANGE Consortium. 2024. Smartchange - horizon europe project. Accessed: 2024-09-02. (2024). https://www.smart-change.eu/.

[4] K. Corder, E. M. van Sluijs, I. Goodyer, C. L. Ridgway, R. M. Steele, D. Bamber, V. Dunn, S. J. Griffin, and U. Ekelund. 2011. Physical activity awareness of british adolescents. *Archives of Pediatrics Adolescent Medicine*, 165, 3, 281–289.

[5] A. García-Hermoso, R. Ramírez-Campillo, and M. Izquierdo. 2019. Is muscular fitness associated with future health benefits in children and adolescents? a systematic review and meta-analysis of longitudinal studies. *Sports Medicine*, 49, 7, (July 2019), 975–989.

[6] D. C. Jr Goff, D. M. Lloyd-Jones, G. Bennett, S. Coady, R. B. D'Agostino, and R. Gibbons. 2014. American college of cardiology/american heart association task force on practice guidelines. 2013 acc/aha guideline on the assessment of cardiovascular risk: a report of the american college of cardiology/american heart association task force on practice guidelines. *Circulation*, 129, 25, (June 2014), S49–S73.

[7] Noelia González-Gálvez, Jose Carlos Ribeiro, and Jorge Mota. 2022. Cardiorespiratory fitness, obesity and physical activity in schoolchildren: the effect of mediation. *International journal of environmental research and public health*, 19, 23, 16262–16270. ObjectType-Article-1. DOI: 10.3390/ijerph192316262.

[8] S. J. Griffin, P. S. Little, C. N. Hales, A. L. Kinmonth, and N. J. Wareham. 2000. Diabetes risk score: towards earlier detection of type 2 diabetes in general practice. *Diabetes/Metabolism Research and Reviews*, 16, 3, 164–171.

[9] D. R. Jacobs, J. G. Woo, A. R. Sinaiko, S. R. Daniels, J. Ikonen, and M. Juonala. 2022. Childhood cardiovascular risk factors and adult cardiovascular events. *New England Journal of Medicine*, 386, 19, (May 2022), 1765–1777.

[10] Markus Juonala et al. 2008. Cohort profile: the cardiovascular risk in young finns study. *International Journal of Epidemiology*, 37, 6, 1220–1226.

[11] Gregor Jurak et al. 2020. Slofit surveillance system of somatic and motor development of children and adolescents: upgrading the slovenian sports educational chart. *Acta Universitatis Carolinae. Kinanthropologica*, 56, 1, 28–40. DOI: 10.14712/23366052.2020.4.

[12] H. C. Jr McGill, C. A. McMahan, E. E. Herderick, G. T. Malcom, R. E. Tracy, and J. P. Strong. 2000. Origin of atherosclerosis in childhood and adolescence. *American Journal of Clinical Nutrition*, 72, 5, (Nov. 2000), 1307S–1315S.

[13] K. Pahkala, H. Hietalampi, T. T. Laitinen, J. S. Viikari, T. Rönnemaa, H. Niinikoski, and et al. 2013. Ideal cardiovascular health in adolescence: effect of lifestyle intervention and association with vascular intima-media thickness and elasticity (the special turku coronary risk factor intervention project for children [strip] study). *Circulation*, 127, 18, (May 2013), 2088–2096.

[14] J. R. Ruiz, I. Cavero-Redondo, F. B. Ortega, G. J. Welk, L. B. Andersen, and V. Martinez-Vizcaino. 2016. Cardiorespiratory fitness cut points to avoid cardiovascular disease risk in children and adolescents; what level of fitness should raise a red flag? a systematic review and meta-analysis. *British Journal of Sports Medicine*, 50, 13, 773–779.

[15] T. J. Saunders, C. E. Gray, V. J. Poitras, J. P. Chaput, I. Janssen, P. T. Katzmarzyk, and et al. 2016. Combinations of physical activity, sedentary behaviour and sleep: relationships with health indicators in school-aged children and youth. *Applied Physiology, Nutrition, and Metabolism*, 41, 6, (June 2016), 486–505.

[16] Johan Simons, Gaston Beunen, Roland Renson, Albrecht L. M. Claessens, Bernard Vanreusel, and Jos A. V. Lefevre. 1990. *Growth and fitness of Flemish girls: The Leuven Growth Study*. Human Kinetics, Champaign, IL.

[17] SCORE2 working group and ESC Cardiovascular risk collaboration. 2021. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *European Heart Journal*, 42, 25, (June 2021), 2439–2454.

[18] World Health Organization. 2018. Global Health Estimate 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. World Health Organization.

Predicting Mental States During VR Sessions Using Sensor Data and Machine Learning

Emilija Kizhevska*
emilija.kizhevska@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School (IPS)
Ljubljana, Slovenia

Mitja Luštrek
mitja.lustrek@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School (IPS)
Ljubljana, Slovenia

Abstract

Empathy is a multifaceted concept with both cognitive and emotional components that plays a crucial role in social interactions, prosocial behavior, and mental health. In our study, empathy and general arousal were induced via VR, with physiological signals measured and ground truth collected through questionnaires. Data from over 100 participants were collected and analyzed using multiple machine learning models and classification algorithms to predict empathy based on physiological responses. We explored different data balancing techniques and labeled data in multiple ways to enhance model performance. Our results show that they are effective in detecting general arousal, empathy, and differentiating between non-empathic and empathic arousal, but the models encountered difficulties with precise emotion detection. The dataset extracted at 5-second intervals and models using Random Forest and Extreme Gradient Boosting showed the best performance. Future work will focus on refining emotion detection through advanced modeling techniques and investigating gender differences in empathy.

Keywords

VR, mental states, machine learning, sensor data

1 Introduction

Empathy is a multifaceted concept explored across various fields, including psychology, neuroscience, and sociology. Though no universal definition exists, empathy is generally understood to include both cognitive (understanding another's perspective) and emotional (experiencing another's feelings) components [8]. Our research defines empathy as the ability to model others' emotional states and respond sensitively while recognizing the self-other distinction [14].

There is no "golden standard" for measuring empathy [10], with methods varying from self-report questionnaires to psychophysiological measures like heart rate and skin conductance. Each method has its pros and cons, often leading to a combination of approaches for a comprehensive assessment. Psychophysiological measures offer objective data but face challenges due to individual variability and non-empathetic factors. Our study addresses these issues by using machine learning to directly measure empathy from physiological signals, offering a novel approach.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.9356

VR creates an immersive environment that enhances empathy by allowing users to experience different perspectives and engage emotionally. VR is effective for empathy training and is referred to as 'the ultimate empathy machine' [1, 11] for various reasons: 1) Immersive Experience: Provides a strong sense of presence, helping users adopt new viewpoints [15]. 2) Perspective-Taking and Emotional Engagement: Simulates realistic scenarios to provoke emotional responses and understanding [19]. 3) Empathy Training: Effective in healthcare, education, and diversity training by challenging preconceptions and deepening emotional insights [16]. 4) Ethical Considerations: Ensures respectful use of VR, balancing immersive experiences with participants' well-being [2].

The objective of this study was to examine how participants' empathy correlates with changes in their physiological metrics, measured using sensors such as inertial measurement unit (IMU), photoplethysmograph (PPG), and electromyography (EMG). Participants were immersed in 360° VR videos featuring actors displaying various emotions (sadness, happiness, anger, and anxiety) and reported their empathetic experiences via brief questionnaires. Using data from these sensors and questionnaires, machine learning models were developed to predict empathy scores based on physiological responses during the VR sessions [9].

2 Materials and Data Collection Process

2.1 Materials and Setup for Empathy Elicitation in VR

To elicit empathy, we immersed participants in a 360° and 3D virtual environment, as VR has proven more effective than methods like 2D videos, workshops, or text-based exercises [8, 13, 17, 20]. We used videos featuring actors expressing four emotions—happiness, sadness, anger, and anxiousness—without additional content to avoid confounding factors [2]. Recognizing the impact of understanding emotional context, an audio narrative version was also created, followed by a corresponding video (50-120 seconds). To ensure gender balance, we recorded videos with two male and two female actors. Five versions were developed: four with narratives (two male, two female) and one non-narrative, where all emotions are portrayed by all actors without accompanying narratives. The non-narrative version allows gradual transitions between emotions, making it suitable for participants of all linguistic backgrounds.

Additionally, a 2-minute forest video ("The Amsterdam Forest in Springtime") was included at the start to establish a relaxed baseline and a roller coaster video ("Official 360 POV - Yukon Striker - Canada's Wonderland") at the end to control for non-empathic arousal. Both videos were sourced from YouTube.

Participants completed trait empathy questionnaires (QCAE) [14] and, after each emotion-specific video, provided feedback

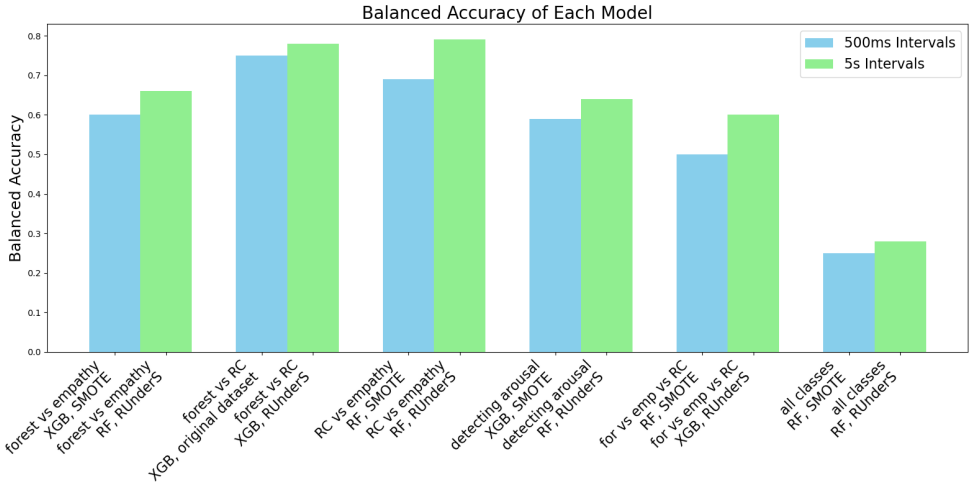


Figure 1: The best accuracies for each group of models, developed using datasets extracted at two different frequencies and various data balancing techniques, presented for all the labeling schemes

on their empathic state (State Empathy Scale) [18], arousal and valence levels (SAM) [3], and personal distress (IRI) [5]. Each VR session lasted around 20 minutes to minimize VR sickness, with participants viewing one of five versions.

Sensor data were collected using the emteqPRO system attached to the Pico Neo 3 Pro Eye VR headset, including EMG for facial muscle activation, PPG for heart rate, and IMU for head motion tracking. The device uses an internal clock as well [12].

2.2 Dataset Description

In this research, we used convenience sampling to recruit participants from the general public without a specific selection pattern. Participants were invited from various sources, including Jožef Stefan Institute employees, university students, and the general public. Invitations were sent verbally or in writing. Data collection concluded with 105 participants, averaging 22.43 ± 5.31 years (range 19–45), with 75.24% identifying as female. Participants had diverse educational and professional backgrounds. Additionally, ethical clearance for this study was obtained from the Research Ethics Committee at the Faculty of Arts, University of Maribor, Slovenia (No. 038-11-146/2023/13FFUM). Furthermore, written informed consent was obtained from the actors prior to recording.

The EmteqPRO system not only provides raw sensor data but also generates derived variables through the Emteq Emotion AI Engine, which utilizes data-fusion and machine learning to analyze multimodal sensor data and assess the user's emotional state. This system provides a file with 29 derived features, called affective insights for each recording: 7 features for heart-rate variability (HRV) and 3 for breathing rate; 2 features for facial expressions; 4 features for arousal and 4 for valence; 1 feature for facial activation; and 1 feature for facial valence. Additionally, head activity is tracked, reflecting the percentage of the session with head movement. Dry EMG electrodes on facial muscles such as the zygomatic, corrugator, frontalis, and orbicularis provide four more features, each representing muscle activation as a percentage of maximum activation observed during calibration. The data also includes the time elapsed since the start of the recording and the row index.

3 Methodology

3.1 Preprocessing

Since all the features or insights are numeric, except for the feature "Expression/Type," which has three values—smile, frown, and neutral—we applied one-hot encoding, a technique used in data preprocessing where categorical (non-numeric) variables are transformed into a numerical format. Each unique value in the original non-numeric feature is transformed into a separate binary (0 or 1) feature.

Next, because missing values represent less than 1% of the total data for each participant, they were filled in using the average of each feature's values. Scaling the values in the descriptive features between 0 and 1 was the final step in the preprocessing process.

3.2 Feature Engineering

Since features were provided at intervals ranging from 1 second to 500 milliseconds, we divided the data into two windows: one of 5 seconds and one of 500 milliseconds. For each window, we computed features from the 22 insights across the seven modules, as well as from the features for head activity and facial muscle electrodes, deriving a total of 108 new features, including minimum, maximum, average, and standard deviation for each original feature or insight. Additionally, the features for head activity and facial muscle electrodes were used to define 'Expression/Type,' and the time and row index were used as provided. However, the row index was disregarded further in the study.

We labeled the dataset in six different ways: 1) as a binary classification aiming to detect empathic arousal, comparing empathic parts with the forest part of the video, while excluding the non-empathic content of the roller coaster video; 2) as a binary classification using the forest and roller coaster, aiming to detect non-empathic arousal; 3) again, as a binary classification, but including only empathic parts and the roller coaster, aiming to distinguish between empathic and non-empathic arousal, and examining the differences in physiological responses between empathic content and non-empathic arousal-inducing content, such as the roller coaster video; 4) aiming to detect arousal in

general, regardless of whether it is empathic or non-empathic, by splitting the entire dataset into two classes: the forest and everything else, including empathic parts and the roller coaster; 5) into three classes: treating the chunks of the roller coaster and forest as separate classes and grouping all the empathic parts into one class, without differentiating between the different emotions. The goal is to distinguish among no-arousal, empathic arousal, and non-empathic arousal; 6) with the average of participants' answers to the state empathy questions for each part of the video, with each part of the empathic content considered a separate chunk. Additionally, there are two other classes: the forest and the roller coaster. The aim is to detect the level of empathy participants experience during the session. We also included each participant's ID, intending to later use it for model evaluation with the 'leave-one-subject-out' technique.

4 Experiments and Results

4.1 Experimental setup

To build models for predicting a participant's state empathy during the VR session, we used six different classification algorithms: Gaussian Naive Bayes, Stochastic Gradient Descent Classifier, K-Nearest Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier, and Extreme Gradient Boosting Classifier. The balanced accuracy score was used as an evaluation metric to assess the classification models for predicting participants' state empathy. This metric evaluates the overall balanced accuracy of the model by calculating the average of recall obtained on each class. Additionally, we used a confusion matrices to evaluate the performance of the classification models by comparing the actual and predicted labels.

For model evaluation, we used a Leave-One-Subject-Out cross-validation setup, where each subject is a unique participant identified by their ID.

Because the labeling schemes 2, 3, 5, and 6 are not balanced (with the 80% of the majority class), we conducted four experiments for each developed model: 1) applying the Synthetic Minority Over-sampling Technique (SMOTE) to create synthetic samples for the minority class to balance the dataset; 2) using the RandomUnderSampler (RUnderS) method to randomly select samples from the majority class, thereby reducing their count and balancing the dataset; 3) using SMOTETomek, a combination of SMOTE for oversampling and Tomek links for undersampling, which targets both the minority and majority classes; and 4) using the dataset as it is, without any undersampling or oversampling.

4.2 Results

Including models developed by six different classification algorithms on two distinct datasets—with two different window sizes—and utilizing four different data balancing techniques: undersampling, oversampling, combination techniques, and the dataset in its original form, along with six different labeling schemes, we obtained 288 unique confusion matrices and corresponding accuracies for each combination.

We ran a correlation matrix, which revealed that the highest correlation with the state empathy feature was found with the derived maximum and minimum values from the mean heart rate, the derived maximum and minimum values from the arousal class feature, and the average of the arousal class — the insight, which can be -1 (low), 0 (medium), or 1 (high). The derived standard deviation, maximum, and minimum values from the activation—expressed as a percentage of the maximum activation of particular muscles from the calibration session, especially the zygomaticus and orbicularis muscles—were also highly correlated.

Regarding the labeling schemes, we can conclude the following: 1) We can detect empathic arousal with confusion matrices that show a relatively good distribution of correct predictions across both classes and high accuracies for most of the developed models; 2) We can detect non-empathic arousal, with almost every developed model achieving a balanced accuracy higher than 60%, reaching up to 78%, and a reasonable balance between classes, indicating satisfactory classification performance; 3) We can even distinguish between empathic and non-empathic arousal with balanced accuracy of 79%; 4) We can detect arousal in general, again with high accuracies and balanced classes; 5) We can distinguish to some extent among no-arousal, empathic arousal, and non-empathic arousal; 6) However, it is currently very challenging to detect the precise level of empathy participants are feeling during the session using these methods, and to determine whether they are empathizing by mirroring emotions or experiencing something different while observing specific emotions. The best we can detect in this regard is up to 28% balanced accuracy, with confusion matrices showing a relatively balanced performance across multiple classes, with a good number of correct classifications, particularly in the more frequent classes.

Regarding the two window sizes, both models showed similar class balance and balanced accuracy scores. However, the dataset extracted at 5-second intervals performed slightly better. Using this dataset, false positives and false negatives were reduced more effectively. This led to more reliable classification performance, especially in terms of precision and recall, despite the smaller scale. Thus, the models developed using the 5-second interval dataset generally performed better, showing more effective classification and fewer errors. The simpler confusion matrix and potentially better handling of fewer classes suggest that it performs better in practical terms (Figure 2, Figure 1).

Regarding the data balancing techniques, the undersampling technique never produced the best results. For the dataset extracted at 500 ms intervals, using the SMOTE oversampling technique and SMOTETomek yielded the best results. For the dataset extracted at 5-second intervals, using the entire dataset yielded the best results, although models developed using SMOTETomek yielded slightly lower results in each combination of different labeling schemes.

Regarding the classification algorithms, Gaussian Naive Bayes performed the worst in terms of balanced confusion matrices, while Random Forest Classifier and Extreme Gradient Boosting performed the best across all combinations, with Random Forest Classifier showing slightly better results for most combinations (Figure 2, Figure 1).

4.3 Conclusion

In this study, we define the entire plan for developing materials, methods, and environments to evoke and measure the level of empathy. We started by defining the videos and the session, creating or selecting questionnaires for later use as ground truth, writing the narratives, recording the VR videos, and then editing and preparing them for use. Additionally, we collected a dataset from over 100 participants, which we filtered, preprocessed, and prepared for feature engineering and analysis.

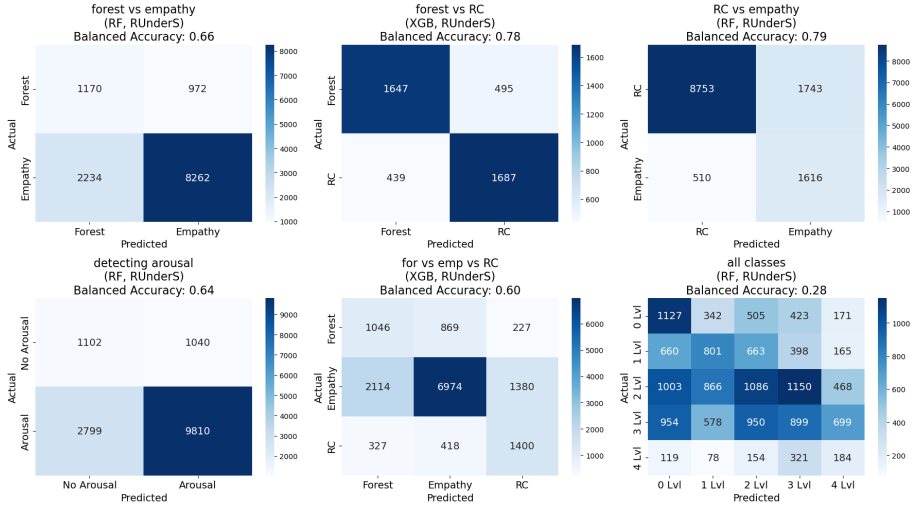


Figure 2: The best confusion matrices for each group of models, developed using dataset extracted at a 5s window size and various data balancing techniques, shown for all labeling schemes

We conducted and analysed four groups of experiments, totaling 288 combinations, where we developed models using two different window sizes, six classification algorithms, and three resampling techniques, with six different labeling schemes aimed at detecting various aspects of the dataset chunks: four empathetic parts, forest, and roller coaster.

The main conclusion is that we can detect arousal in general, non-empathic arousal, empathy, and differentiate between non-empathic and empathic arousal, as well as between relaxed states and arousal. However, we face difficulties in detecting and distinguishing between the precise levels of empathy during VR sessions using these methods and approaches.

Our next steps involve refining the detection of empathy levels during VR sessions by applying detailed data filtering and transforming it into a stationary format. Furthermore, we will develop models such as Autoregressive, Moving Average, and Extended Recurrent Moving Average, and use clustering techniques like DBSCAN and HDBSCAN. Additionally, we will extract more features from the raw data or use end-to-end neural networks. We plan to analyze gender differences in empathy with a t-test [7], and explore the impact of narrative context and emotions on empathic responses using ANOVA and MANOVA [4, 6].

Acknowledgements

The part of Emilija Kizhevska was supported by the Slovenian Research and Innovation Agency (ARIS) as part of the young researcher PhD program, grant PR-12879. The technical aspects of the videos, the recording and video editing were skillfully conducted by Igor Djilas and Luka Komar. The actors featured in the videos were Sara Janašković, Kristýna Šajtošová, Domen Puš, and Jure Žavbi. The questionnaires were selected and created, the narratives were written, and the psychological aspects of the video creation were considered by Kristina Šparemblek.

References

[1] D. Banakou, P. D. Hanumanthu, and M. Slater. 2016. Virtual embodiment of white people in a black virtual body leads to a sustained reduction in their implicit racial bias. *Frontiers in Human Neuroscience*, 10, 226766.

[2] P. Bertrand, J. Guegan, L. Robieux, C. A. McCall, and F. Zenasni. 2018. Learning empathy through virtual reality: multiple strategies for training empathy-related abilities using body ownership illusions in embodied virtual reality. *Frontiers in Robotics and AI*, 5, 326671.

[3] M. M. Bradley and P. J. Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25, 1, 49–59.

[4] A. Cuevas, M. Febrero, and R. Fraiman. 2004. An anova test for functional data. *Computational Statistics and Data Analysis*, 47, 1, 111–122.

[5] M. H. Davis. 1980. A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology/American Psychological Association*, 85.

[6] A. French, M. Macedo, J. Poulsen, T. Waterson, and A. Yu. 2008. Multivariate analysis of variance (manova). *San Francisco State University*.

[7] T. K. Kim. 2015. T test as a parametric statistic. *Korean Journal of Anesthesiology*, 68, 6, 540–546.

[8] E. Kizhevska, F. Ferreira-Brito, T. Guerreiro, and M. Luštrek. 2022. Using virtual reality to elicit empathy: a narrative review. *VR4Health@ MUM*, 19–22.

[9] E. Kizhevska, K. Šparemblek, and M. Luštrek. 2024. Protocol of the study for predicting empathy during vr sessions using sensor data and machine learning. *PloS One*, 19, 7, e0307385.

[10] F. F. D. Lima and F. D. L. Osório. 2011. Empathy: assessment instruments and psychometric quality–a systematic literature review with a meta-analysis of the past ten years. *Frontiers in Psychology*, 12. 781346.

[11] M. Mado, F. Herrera, K. Nowak, and J. Bailenson. 2021. Effect of virtual reality perspective-taking on related and unrelated contexts. *Cyberpsychology, Behavior, and Social Networking*, 24, 12, 839–845.

[12] M. J. Magnée, B. De Gelder, H. Van Engeland, and C. Kemner. 2007. Facial electromyographic responses to emotional information from faces and voices in individuals with pervasive developmental disorder. *Journal of Child Psychology and Psychiatry*, 48, 11, 1122–1130.

[13] K. M. Nelson, E. Anggraini, and A. Schlüter. 2020. Virtual reality as a tool for environmental conservation and fundraising. *PloS One*, 15, 4, e0223631.

[14] R. L. Reniers, R. Corcoran, R. Shryane Drake, N. M., and B. A. Völlm. 2011. The qcae: a questionnaire of cognitive and affective empathy. *Journal of personality assessment*, 93, 1, 84–95. DOI: doi:10.1080/00223891.2010.528484.

[15] G. Riva, J. A. Waterworth, and E. L. Waterworth. 2004. The layers of presence: a bio-cultural approach to understanding presence in natural and mediated environments. *CyberPsychology Behavior*, 7, 4, 402–416.

[16] R. O. Roswell, C. D. Cogburn, J. Tocco, J. Martinez, C. Bangeranye, J. N. Bailenson, and L. Smith. 2020. Cultivating empathy through virtual reality: advancing conversations about racism, inequity, and climate in medicine. *Academic Medicine*, 95, 12, 1882–1886.

[17] N. S. Schutte and E. J. Stilinović. 2017. Facilitating empathy through virtual reality. *Motivation and Emotion*, 41, 708–712.

[18] L. Shen. 2010. On a scale of state empathy during message processing. *Western Journal of Communication*, 74, 5, 504–524.

[19] M. Slater, A. Antley, A. Davison, D. Swapp, C. Guger, C. Barker, and M. V. Sanchez-Vives. 2006. A virtual reprise of the stanley milgram obedience experiments. *PloS One*, 1, 1, e39. DOI: 10.1145/1188913.1188915.

[20] J. Stargatt, S. Bhar, T. Petrovich, J. Bhowmik, D. Sykes, and K. Burns. 2021. The effects of virtual reality-based education on empathy and understanding of the physical environment for dementia care workers in australia: a controlled study. *Journal of Alzheimer's Disease*, 84, 3, 1247–1257.

Biomarker Prediction in Colorectal Cancer Using Multiple Instance Learning

Miljana Shulajkovska*
miljana.sulajkovska@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Matej Jelenc
jelenc11matej@gmail.com
Jožef Stefan Institute
Ljubljana, Slovenia

Jitenndra Jonnagaddala
jitendra.jonnagaddala@unsw.edu.au
School of Population Health, Faculty of Medicine and Health
Syndey, Australia

Anton Gradišek
anton.gradisek@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

Microsatellite instability (MSI) is a crucial biomarker in colorectal cancer, guiding personalised treatment strategies. The focus of our paper is on evaluating how different state-of-the-art pretrained artificial intelligence models perform in extracting features on molecular and cellular oncology (MCO) study dataset to predict biomarkers. In this study, we present an advanced approach for MSI prediction using multiple instance learning on whole slide images. Our process begins with comprehensive preprocessing of WSIs, followed by tessellation, which breaks down large images into manageable tiles. State-of-the-art feature extraction techniques are utilised on these selected tiles, employing pretrained models to capture rich, discriminative features. Various aggregation methods are applied to combine these features, leading to the prediction of MSI status across the entire slide. We assess the performance of different pretrained models within this framework, demonstrating their effectiveness in accurately predicting MSI, with results showing an AUROC of 0.91 on the MCO dataset. Our findings underscore the potential of multiple instance learning-based approaches in enhancing biomarker prediction in colorectal cancer, contributing to more targeted and effective treatment strategies.

Keywords

multiple instance learning, whole slide images, colorectal cancer, biomarker prediction

1 Introduction

MSI is a crucial biomarker in colorectal cancer (CRC) that indicates defects in the DNA mismatch repair system, leading to a high mutation rate within tumor cells. MSI status has significant clinical implications, influencing treatment decisions, particularly the use of immunotherapy, and providing prognostic information. Traditionally, MSI is determined through laboratory tests such as PCR-based assays or immunohistochemistry (IHC) on tumor tissue samples, which require invasive biopsy procedures. However, these methods can be time-consuming, costly, and dependent on the availability of sufficient tissue samples.

Deep learning methods have emerged as a promising non-invasive alternative for MSI prediction by analysing whole slide images (WSIs) of histopathological samples. These models can detect patterns linked to MSI, eliminating the need for genetic testing. WSIs provide a comprehensive view of tumor histology, offering a faster, less invasive, and more accessible means of diagnosis.

Integrating deep learning into clinical practice can improve early MSI detection, personalise treatment, and reduce invasive procedures. WSI-based methods streamline diagnostics and enhance cancer care with accessible predictive analytics.

To manage these challenges, WSIs are often divided into smaller regions or patches. A common method to address these issues is Multiple Instance Learning (MIL) [3, 8]. Due to the vast size of WSIs, computational resources can be easily overwhelmed, making MIL an essential approach. MIL is a machine learning technique that operates on sets or "bags" of instances, where the label is assigned to the entire bag rather than individual instances. This is particularly advantageous in WSI analysis, where labels such as MSI status apply to the entire slide, which is composed of numerous smaller regions or patches.

In this context, [4] demonstrates state-of-the-art (SOTA) results in predicting MSI in colorectal cancer. Their workflow utilizes the Swin-T model on small datasets to predict MSI. First, a pretrained tissue classification model is employed to filter out non-tissue patches, followed by fine-tuning a pretrained model to classify the remaining patches. Both intra-cohort and external validation are performed. When trained on the MCO dataset (N=1065), the model achieved a mean AUROC of 0.92 ± 0.05 for MSI prediction. Similarly, [11] employs a transformer-based approach for large-scale multi-cohort evaluation, involving over 13,000 patients for biomarker prediction, achieving a negative predictive value of over 0.99 for MSI prediction. When trained and tested only on a single cohort (MCO), the model achieved an AUROC of 0.85. While [4] achieved promising results on the MCO dataset using an additional tissue classifier, we obtained comparable performance without the need for tissue classification. On the other hand, [11] used a multicentric cohort, which demands additional computational resources. In comparison to their results on the MCO dataset, we achieved a 6% improvement using a smaller dataset.

In this study, we leverage MIL to process WSIs for the prediction of MSI in CRC. By testing SOTA models on the MCO dataset, we aim to assess their performance in MSI prediction using MIL. This approach not only highlights the potential of MIL in processing complex, unannotated WSIs but also contributes

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.9705

to the broader goal of improving biomarker prediction in CRC, ultimately supporting more personalized and effective treatment strategies.

The paper is organised as follows: Section 2 outlines the methods used in the pipeline, Section 3 provides a description of the data, Section 4 presents the results, and Section 5 discusses the findings and potential directions for future work.

2 Methods

This section outlines the pipeline for MSI prediction, as illustrated in Figure 1. The process begins with the preprocessing of WSIs, including tessellation into smaller patches. Next, SOTA pretrained models are employed to extract features from these patches. These models, trained on large and diverse datasets, capture rich and discriminative features crucial for accurate MSI prediction. Finally, aggregation techniques are applied to combine the information from the patches, enabling precise MSI status prediction for the entire slide. Each subsection provides a concise explanation of these individual processes.

2.1 Preprocessing

WSIs are first tessellated into smaller, more manageable patches to facilitate further processing. This step involves dividing the large images into smaller regions using the tiatoolbox presented in [9]. Non-informative tissue patches are removed to ensure the analysis focuses solely on relevant tissue areas.

Specifically, patches that are out of bounds—where only a portion contains actual image data and the remainder consists of padding—are discarded. Patches that consist entirely of tissue are retained for subsequent analysis. This preprocessing step ensures that only informative and relevant patches are used for feature extraction and MSI prediction.

2.2 Feature Extraction Methods

Since only WSI-level annotations are available, several pretrained feature extraction models - UNI [1], ProvGigaPath [13], Phikon [2] and CTransPath [12] - are applied to patches, removing the need for detailed patch-level labeling. These SOTA models, trained on large datasets, can capture complex and discriminative features essential for accurate biomarker prediction. The extracted feature embeddings are then used as input for the aggregation and classification stages, laying the foundation for precise MSI status prediction. For technical details about these models, see Table 1.

2.3 Aggregation Methods

After feature extraction, we apply aggregation techniques to combine patch-level features into a slide-level representation. Traditional pooling methods like max-pooling and mean-pooling provide straightforward approaches.

However, these methods are limited by their lack of trainability. In recent years, attention-based pooling or ABMIL became a popular technique that adresses this issue [6]. ABMIL assigns a weight α_i to each patch's feature vector, reflecting its importance:

$$F = \sum_{i \in P} \alpha_i f_i$$

The attention scores α_i are computed as:

$$\alpha_i = \frac{\exp(w^\top \tanh(V f_i))}{\sum_{k \in P} \exp(w^\top \tanh(V f_k))}$$

where w and V are trainable parameters.

This approach allows the model to dynamically focus on the most relevant patches, leading to more accurate MSI predictions.

Another technique similiar to attention is DSMIL [7] or a dual stream aggregator, consisting of two branches, employing both an instance classifier and a bag classifier. Let $h_i \in \mathbb{R}^{L \times 1}$ be a feature embedding, and $B = \{h_0, ..., h_n\}$ a bag of embeddings. The first stream uses an instance classifier, followed by a max-pooling operation to obtain a score $c_m(B)$ and the critical embedding h_m. The second stream aggregates the embeddings into a single bag embedding which is then passed through a bag classifier:

$$c_b(B) = W_b \sum_{i}^{n-1} U(h_i, h_m) v_i$$

Where W_b is a weight vector for classification, v_i an information vector and U is a distance measurement between an arbitrary embedding and the critical embedding:

$$U(h_i, h_m) = \frac{\exp(\langle q_i, q_m \rangle)}{\sum_{k=0}^{n=1} \exp(\langle q_k, q_m \rangle)}$$

where is a query vector. Both q_i and v_i are calculated by:

$$q_i = W_q h_i, \quad v_i = W_v h_i, \quad i = 0, ..., n-1$$

where W_q and W_v are weight matrices. The final prediction is given by:

$$c(B) = \frac{1}{2}(c_m(B) + c_b(B))$$

The last approach for feature aggregation reviewed in this paper is TransMIL, as proposed in [10], a Transformer based aggregation method, which unlike the afore-mentioned methods, takes into account spatial information as well. By treating a bag of embeddings as a sequence of tokens, TransMIL uses a novel TPT module made up of two Transformer layers and a position encoding layer, where Transformer layers are designed for aggregating morphological information and Pyramid Position Encoding Generator (PPEG) which encodes spatial information, followed by a multi-layer perceptron (MLP) which classifies the bag.

2.4 MSI Classification

The aggregation step produces a single feature vector F, which encapsulates the most informative characteristics of the entire slide. This aggregated feature vector F is then passed through one or more fully connected (dense) layers. These layers apply learned weights and biases to the features to transform them into a form that is more suitable for classification. The output of the fully connected layer is often passed through an activation function, such as a sigmoid or softmax, depending on whether the classification task is binary (microsatellite instability MSI vs. microsatellite stability MSS) or multi-class. For MSI prediction, a sigmoid function is typically used, outputting a probability value between 0 and 1. The final output of the model is a single probability value indicating the likelihood of the slide being MSI. A threshold (e.g., 0.5) is applied to this probability to make a binary decision.

3 Data

For this paper the MCO study [5] was used for training and testing. The MCO study collection contains 1,500 digitized whole slide images (WSIs) of colorectal cancer tissues. Conducted by the Molecular and Cellular Oncology (MCO) Study group from 1994 to 2010, this study systematically gathered tissue samples

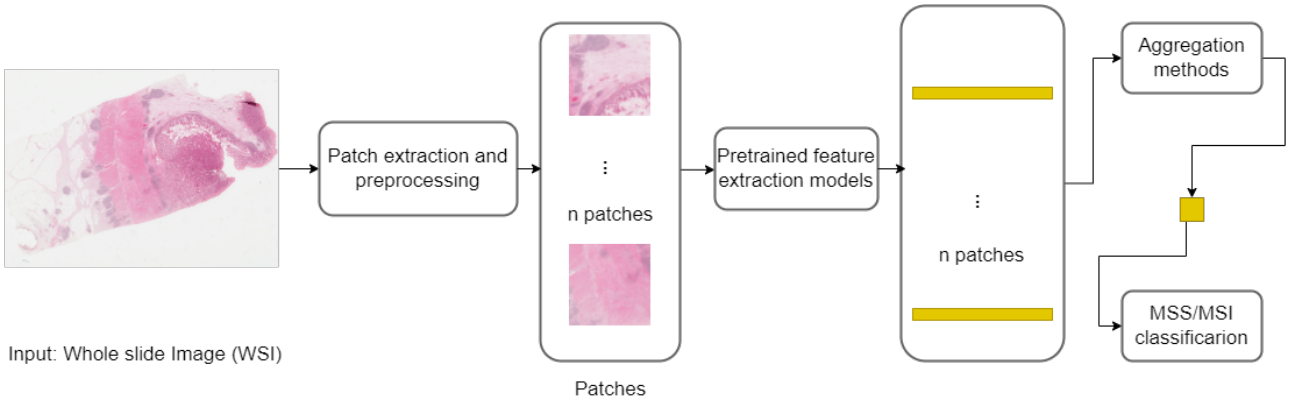


Figure 1: General architecture: multiple-instance learning approach.

feature extractor	architecture	dataset	embedding size
UNI [1]	ViT-large, DINOv2, 16 heads	Mass-100k: in-house histopathology slides from MGH and BWH, and external slides from the GTEx consortium containing >100M images, derived from >100,000 WSIs across 20 major tissue types	1024
ProvGigaPath [13]	ViT-large, DINOv2, 24 heads	Prov-Path: dataset from Providence, a large US health network comprising 28 cancer centres, consisting of 1,3B images from 171,189 WSIs	1536
Phikon [2]	ViT-large, iBOT combining MIM and CL	PanCancer40M: dataset from TCGA, covering 13 anatomic sites and 16 cancer subtypes, consisting of 43,4M images from 6,093 WSIs	768
CTransPath [12]	CNN with multi-scale Swin Transformer	dataset from TCGA and PAIP, consisting of 15M images from 32,220 WSIs	768

Table 1: Technical details about the pretrained feature extraction models.

and clinical data from over 1,500 patients who underwent colorectal cancer surgery. Each slide, representing a typical tumor section, is stained with Hematoxylin and eosin and scanned at a 40x objective, achieving a resolution of 0.25 mpp comparable to an optical microscope (~100,000 dpi). The total data size is approximately 3 Terabytes, and the collection is available on the Intersect Australia RDSI Node.

4 Results

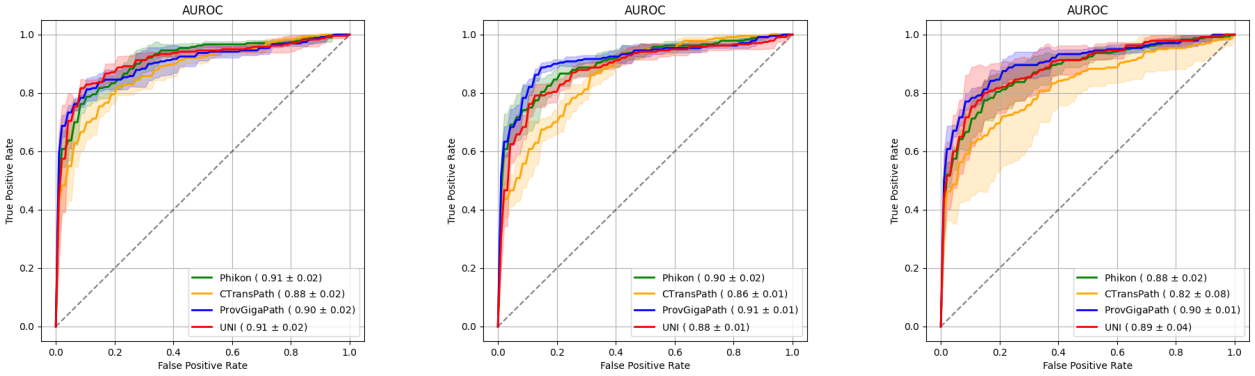
The dataset used in this study comprised 996 whole slide images (WSIs), with 242 labeled as MSI and 754 as MSS. To evaluate the performance of various aggregation methods, models were trained using 5-fold cross-validation, which ensured robust training and validation. To create a balanced testing set of 96 samples, 20% of positive (MSI) samples and an equal number of negative (MSS) samples were randomly excluded. The remaining data was split into five equally balanced parts for cross-validation, with each fold consisting of 180 samples in the validation set and 720 samples in the training set.

WSIs were then preprocessed into bags, each containing approximately 2,000 to 4,000 patches. Each patch was then converted into feature embeddings using four different feature extraction methods: Phikon, CTransPath, ProvGigaPath, and UNI. Specifically, CTransPath and Phikon produced embeddings with 768 features, UNI with 1024 features, and ProvGigaPath with 1536 features.

Three feature aggregation methods—ABMIL, DSMIL, and TransMIL—were applied to the extracted features to generate a single representative feature for each WSI. Following aggregation, a simple neural network with a sigmoid activation function and a threshold of 0.5 was used to classify MSI and MSS.

Each aggregation model was then trained for each feature extraction method on each fold, with training being conducted over 50 epochs using the AdamW optimiser and the 1-cycle learning rate scheduler to adjust the learning rate as models approached convergence. Binary cross-entropy (BCE) was used as the loss function. After each epoch, model performance was evaluated on the validation set using the AUROC metric to select the best checkpoint, as most models tended to overfit toward the end of training. The selected checkpoints were then tested to calculate the mean AUROC across all folds.

Results are presented in Figure 2a. The best performance was achieved using the DSMIL aggregation method with the ProvGigaPath feature extractor, yielding an AUROC of 0.91 ± 0.01. The ABMIL method performed best with the Phikon and UNI extractors, achieving AUROCs of 0.91 ± 0.02. Finally, the TransMIL method combined with ProvGigaPath resulted in an AUROC of 0.90 ± 0.01. Additionally, statistical analysis was performed, specifically, the Wilcoxon signed-rank test, which yielded an average p-value of 0.446, showing a relatively insignificant difference in performance of different feature extraction methods, as expected.



(a) ABMIL **(b) DSMIL** **(c) TransMIL**

Figure 2: Predictive performance of 5-fold cross-validation of different feature extractors and aggregation methods. AUROC plots for prediction of MSI/MSS status. The true positive rate represents sensitivity and the false negative rate represents 1-specificity. The shaded areas represent the standard deviation (SD). The value of the lower right each plot represents mean AUROC ± SD.

5 Discussion and Conclusion

n this study, we explored the potential of MIL combined with SOTA pretrained models for predicting MSI in colorectal cancer. Our results indicate that the approach is highly effective, achieving an AUROC of 0.913 on the MCO dataset. This is a notable achievement, particularly when compared to previous studies, such as [4] and[11], which reported AUROCs of 0.92 and 0.85, respectively, on the same dataset. Our results not only validate the effectiveness of our approach but also suggest that the careful selection and combination of feature extraction and aggregation methods can yield improvements in predictive accuracy.

The positive and negative rates observed in our results reflect the model's ability to correctly classify MSI and MSS cases. A high true positive rate (sensitivity) indicates the model's proficiency in identifying MSI-positive cases, which is crucial for ensuring that patients who could benefit from MSI-targeted therapies are accurately identified. Conversely, a high true negative rate (specificity) shows the model's effectiveness in correctly classifying MSS cases, thereby minimising false positives. To further enhance the accuracy and reliability of MSI prediction, several avenues for future work are planned.

Utilisation of the Entire Dataset: We plan to leverage the full dataset to improve the robustness of our model. Training on a larger dataset may help in capturing more nuanced patterns and variations, leading to even more accurate predictions.

Fine-Tuning of Pretrained Models: While we used pretrained models without fine-tuning in this study, fine-tuning these models specifically for the task of MSI prediction could further improve their performance. Tailoring the models to our specific data distribution and task requirements may yield significant gains in accuracy.

Incorporation of a Tissue Classifier: Since MSI is typically found in tumor tissue, we plan to integrate a tissue classifier to automatically remove non-tumor tissue from the analysis. This step should enhance the model's focus on relevant tissue regions, potentially improving MSI prediction accuracy and speed up the whole process.

Development of Advanced Aggregation Methods: We also plan to explore more sophisticated aggregation techniques that can better capture the complex relationships between patches within a WSI. Advanced methods may help in refining the prediction process, leading to further improvements in model performance.

Overall, our study demonstrates the potential of MIL-based approaches in enhancing biomarker prediction in colorectal cancer, paving the way for more personalized and effective treatment strategies.

References

[1] Richard J Chen et al. 2024. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30, 3, 850–862.

[2] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. 2023. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023–07.

[3] Michael Gadermayr and Maximilian Tschuchnig. 2024. Multiple instance learning for digital pathology: a review of the state-of-the-art, limitations & future potential. *Computerized Medical Imaging and Graphics*, 102337.

[4] Bangwei Guo, Xingyu Li, Jitendra Jonnagaddala, Hong Zhang, and Xu Steven Xu. 2022. Predicting microsatellite instability and key biomarkers in colorectal cancer from h&e-stained images: achieving sota predictive performance with fewer data using swin transformer. *arXiv preprint arXiv:2208.10495*.

[5] Nick Hawkins. 2015. MCO study whole slide image collection. (2015).

[6] Maximilian Ilse, Jakub Tomczak, and Max Welling. 2018. Attention-based deep multiple instance learning. In *International conference on machine learning*. PMLR, 2127–2136.

[7] Bin Li, Yin Li, and Kevin W Eliceiri. 2021. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14318–14328.

[8] Oded Maron and Tomás Lozano-Pérez. 1997. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10.

[9] Johnathan Pocock et al. 2022. Tiatoolbox as an end-to-end library for advanced tissue image analytics. *Communications medicine*, 2, 1, 120.

[10] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. 2021. Transmil: transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34, 2136–2147.

[11] Sophia J Wagner et al. 2023. Transformer-based biomarker prediction from colorectal cancer histology: a large-scale multicentric study. *Cancer Cell*, 41, 9, 1650–1661.

[12] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. 2022. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81, 102559.

[13] Hanwen Xu et al. 2024. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 1–8.

Feature-Based Emotion Classification Using Eye-Tracking Data

Tomi Božak
tb85088@student.uni-lj.si
Jožef Stefan Institute
Ljubljana, Slovenia

Mitja Luštrek
mitja.lutsrek@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Gašper Slapničar
gasper.slapnicar@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

The field of emotion recognition from eye-tracking data is well-established and offers near-real-time insights into human affective states. It is less obtrusive than some other modalities, such as electroencephalogram (EEG), electrocardiogram (ECG) and galvanic skin response (GSR), which are often used in emotion recognition tasks. This study examined the practical feasibility of emotion recognition using an eye-tracker with a lower frequency than that typically employed in similar research. Using ocular features, we explored the efficacy of classical machine learning (ML) models in classifying four emotions (anger, disgust, sadness, and tenderness) as well as neutral and "undefined" emotions. The features included gaze direction, pupil size, saccadic movements, fixations, and blink data. The data from the "emotional State Estimation based on Eye-tracking database" was preprocessed and segmented into various time windows, with 22 features extracted for model training. Feature importance analysis revealed that pupil size and fixation duration were most important for emotion classification. The efficacy of different window lengths (1 to 10 seconds) was evaluated using Leave-One-Subject-Out (LOSO) and 10-fold cross-validation (CV). The results demonstrated that accuracies of up to 0.76 could be achieved with 10-fold CV when differentiating between positive, negative, and neutral emotions. The analysis of model performance across different window lengths revealed that longer time windows generally resulted in improved model performance. When the data was split using a marginally personalised 10-fold CV within video, the Random Forest Classifier (RF) achieved an accuracy of 0.60 in differentiating between the six aforementioned emotions. Some challenges remain, particularly in regard to data granularity, model generalization across subjects and the impact of downsampling on feature dynamics.

Keywords

eye-tracking, emotion recognition, machine learning

1 Introduction

Emotion recognition is a vibrant area of research, leveraging diverse data sources such as images [11], audio [16], and also, ocular features like pupil dilation, gaze direction, blinks, and saccadic movements [3, 8, 12]. Such eye-related features provide valuable insights into emotional states, offering a less-invasive and real-time approach to understanding human affective responses. Most studies that tried to predict emotions from these eye-related features relied not only on eye-tracking data but also on EEG [8, 12]. We hypothesized that eye-tracking data is a valuable modality for multi-modal emotion recognition on its own, with potential applications in real-world scenarios like office work, driving, and psychological assessments, as well as in estimating well-being. Our motivation was to explore eye-tracker-based predictive models as an essential component in such practical applications.

The primary objective of our study was to validate existing findings on the performance of classical ML models for emotion classification from eye-tracking data, using the models – Support Vector Machine (SVM) and k-Nearest Neighbors (KNN) – and features already explored in the literature [9, 15] as well as exploring classifiers not so frequently used in this field – such as RF and XGBoost (XGB). Additionally, we aimed to explore the potential of emotion recognition at lower sampling frequencies available in most non-professional eye trackers. For the early feasibility study, we used an existing dataset, which collected data using a wearable eye-tracker but findings could possibly be extended to high-quality unobtrusive contact-free trackers. Our research also focused on understanding the impact of individual features and window lengths on model performance.

2 Related Work

In literature, various physiological signals have been employed for emotion recognition, with a particular focus on modalities such as EEG, GSR, and eye-tracking systems [1, 6, 9]. Researchers have explored both uni- and multi-modal approaches, finding that the integration of multiple modalities can significantly enhance emotion recognition accuracy. Lu et al. achieved 0.78 accuracy with eye-related features recorded with eye-tracking glasses – which are not contact-free but record at relatively low frequencies of 60 Hz or 120 Hz. They predicted positive, negative and neutral classes with SVM. Interestingly, they observed a 0.10 increase in accuracy when combining eye-related and EEG features [12]. Similarly, Guo et al. observed a more substantial gain, with accuracy improving by 0.20 when integrating EEG, eye-tracking, and eye images, as opposed to using only eye-tracking data [7].

The features derived from eye-tracking have been widely used in ML algorithms to detect emotional states [2, 7, 12, 15]. However, most studies have traditionally categorized emotions into broad groups like positive, negative, and neutral [12, 14]. Pupil size, in particular, has emerged as a valuable indicator for distinguishing between positive and negative emotions [2, 7, 12] . Recent efforts have begun to refine these broad categories, identifying more specific emotions like happiness, sadness, fear, anger, etc. [2, 7, 15]. Although current methods can effectively identify certain emotions such as sadness and fear, further research is needed to reliably differentiate between others like disgust, joy, and surprise [2].

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.9988

3 Methodology

3.1 Data

In our research, we used the "emotional State Estimation based on Eye-tracking database" (eSEEd) [13]. The eSEEd comprises data from 48 participants, each of whom watched 10 carefully selected videos intended to evoke specific emotional responses. After viewing each video, participants ranked their emotions – anger, disgust, sadness, and tenderness – on a scale from 0 to 10. Tenderness, however, is not regarded as one of the basic emotions, but it has been widely utilized in emotion research in recent years [13]. Since the participants had ranked all four emotions for every video, a labelling problem emerged when multiple emotions shared the highest score, in our case, leading to "undefined" labels. In our study, emotions were mapped by applying a set of extraction rules in the following order: if the highest-ranked emotion is below four, the response is labelled as neutral; if multiple emotions share the highest rank, the label is undefined; otherwise, the emotion with the highest rank is chosen. The boundary of four was chosen because the original study on eSEEd constructed this rule and we adapted it from there [13]. Although the initial study design aimed for an even distribution of emotions, neutral responses dominate, representing about one-fourth of the labels (depending on window length).

3.1.1 Data Preprocessing. We have preprocessed the data to make it more suitable for our future research and to reduce its size. We wanted to study the performance of data with a relatively low frequency rate of 60 Hz, which is used by relatively affordable mid-tier eye-trackers, like Tobii Pro Spark. Firstly, the features that were uninformative or could be misleading (e.g. raw tracker signal and timestamps) were removed, and the following set of features was preserved: 2D screen coordinates of gaze points (for standard deviation (std) of screen gaze coordinates), 3D coordinates of gaze points (exclusively for saccade calculations), pupil sizes (a and b of the pupil ellipse), and eye IDs (each eye has its own pupil size features). Secondly, rows containing any NaN values were removed, as there were no large consecutive blocks of such rows and downsampling of the data was planned. Finally, we further downsampled the data to 60 Hz, matching the sampling frequency of a mid-tier eye-tracker. However, we acknowledge that downsampling might lead to the loss of high-frequency information, which could be important for capturing subtle dynamics in gaze behaviour and pupil responses. This is particularly relevant considering that recent studies, such as those by Collins et al. [3] and the SEED project [4, 17], have utilized data collected at much higher frequencies to preserve these subtle dynamics. Therefore, while downsampling makes the data more meaningful to our research and more computationally manageable, it is important to keep in mind the reduced temporal resolution when discussing the results.

Following the preprocessing, window segmentation was applied to the data. This step is essential for analyzing temporal patterns within the data, as it allows for the capture of trends and behaviours over specific time intervals. By segmenting the data into windows, we can improve the robustness of feature extraction and model training, enabling the detection of meaningful patterns that might be obscured in raw, unsegmented data. Additionally, with window segmentation, the number of training instances increases which is commonly better for learning more robust ML models and conducting rigorous evaluation. Hence, multiple window lengths were examined, namely: 1, 3, 5 and 10 s. We used 50% sliding window overlap. From each window, we computed 22 features, belonging to the following groups:

(1) gaze coordinates on screen: std of x and y coordinates
(2) pupil ellipse sizes of a and b for each eye: mean, std
(3) blinks: number; mean and std of duration (all 0 if no blinks)
(4) saccades: number; mean speed; mean, std, total duration
(5) fixations: number; mean, std, total duration

Saccade and, implicitly, fixation calculations were done using existing code based on the algorithm proposed by Engbert et al. [5, 10]. The algorithm calculates the velocity and acceleration of eye movements by using a velocity threshold identification method to detect saccades based on continuous 3D gaze data. In our study we define fixation (interval) as an absence of a saccade (interval), thus one fixation is declared between every two saccades (and before the first and after the last one).

As mentioned previously, our data was imbalanced in terms of class distribution, namely the distributions for anger, disgust, sadness, neutral, tenderness and undefined were 8.7%, 13.6%, 17.5%, 25.7%, 15.8% and 18.7%, respectively. Notably, for the 1 s window length, the number of windows was 67,181, whilst for the 10 s window, the number of instances decreased to 6,507.

3.2 Experiments

We initially examined feature correlation matrices to identify potential correlations between features, as well as between feature and class. Then, we compared the following classifiers from the Scikit-learn library: Random Forest (RF), Support Vector Machines (SVM), k-Nearest Neighbors (kNN), and XGBoost (XGB) from the XGBoost library, as well as an ensemble method majority vote of the aforementioned classifiers. We compared all results against a baseline majority classifier. Each model was trained and tested using its default hyperparameters. To evaluate the models' performance, we implemented multiple CV techniques.

The first CV technique was Leave-One-Subject-Out (LOSO). Secondly, we implemented a marginally personalised 10-fold CV "within video." In this approach, a standard 10-fold CV was performed where 90% of temporally sequential windows were used for training and 10% for testing. The splits were done separately for each video within every subject. All the training data from every video was combined to train a single model, and all the test data was combined to evaluate the model, ensuring that the model was exposed to data from all subjects and videos. We named the experiment "marginally" personalised because most training data does not come from any single subject and is thus not very personalised. Finally, we explored a completely personalised 10-fold CV "within subject." Here, training and testing were done only on data of one subject. In all three CV methods, the instances were never shuffled to preserve temporal and subject sequential information and to minimize overfitting.

We attempted to merge certain classes in a way to group negative emotions – anger, disgust, and sadness – under the category "negative," while labelling tenderness as "positive." The label for neutral remained unchanged, while the undefined label was changed to "negative" because it always resulted from multiple negative emotions scoring equally. Lastly, the feature importances were analysed for different combinations of data splits and models in order to identify potential consistently important features.

4 Results

The results described in the following subsections are summarised in Table 1.

4.1 Feature Correlations

The first important observation from correlation matrices was that no output class is closely correlated to any other singular feature. Secondly, we noticed some strong correlations, for example, a 1.0 correlation between a number of fixations and a number of saccades, because one simply equals the other increased by one. More importantly, we noticed little-to-no correlation between features that proved to be most important in some best-performing models, meaning each of these features brought some novel information to the model. The only exceptions of important features being correlated are the features representing the mean size of a pupil i.e., ellipse a and b axes, which are expected to be correlated. They were correlated more than 0.8. However, we decided not to remove any features because we assessed the feature count of 22 to be well-balanced in relation to the number of instances.

4.2 Leave-One-Subject-Out

With the goal of training a robust general model for our dataset, we first applied the LOSO CV technique. The best performance was achieved by RF on 10 s windows, yielding an accuracy of 0.28 ± 0.13 and an F1-score of 0.28 ± 0.16. It outperformed the majority classifier by 0.03 in accuracy and 0.13 in F1-score. In a subsequent experiment, the negative emotions were grouped. This adjustment led to an overall increase in performance. However, with such grouping the majority classifier score also increased to 0.59 accuracy, which is the same as the best-performing model.

Further analysis revealed that high accuracy mainly implied the subject predominantly reported "neutral" feelings and low accuracy implied little-to-no "neutral" labels. However, not every subject with a high "neutral" count achieved outstanding results and not every subject with a wide range of emotions yielded poor results. A comparison was made between the number of windows in the left-out subject to their performance and no correlation was found. 10 s window length performed better than the shorter windows with lengths 1-5 s. We also tested longer (60 s) windows and the resulting accuracies were higher than those from 10 s windows, but we evaluated that the number of instances was insufficient for the results to be representative.

4.3 Marginally Personalised 10-fold Cross-Validation Within Video

Given that the LOSO yielded relatively poor results, the next step was to explore 10-fold CV. Experiments showed an average accuracy of 0.60 ± 0.07 and an F1-score of 0.60 ± 0.08, produced with RF on 10 s windows, the best-performing model. This should be compared to the results given by the majority classifier – average accuracy of 0.21 ± 0.01 and F1-score of 0.07 ± 0.01. With negative emotions grouped, the accuracy and F1-score raised to 0.76 ± 0.04 and 0.73 ± 0.04, respectively, for the best-performing XGB on 10 s windows. The majority class classifier yielded an accuracy of 0.66 ± 0.02 and an F1-score of 0.52 ± 0.02.

4.4 Personalised 10-fold Cross-Validation

Even though 10-fold CV within video resulted in much better performance compared to LOSO, we wanted to see the performance of completely personalised models. All the models performed similarly well, with the absolute best being RF on 10 s windows which outperformed the majority classifier by 0.05 and 0.13 for accuracy and F1-score, respectively. When grouping the negative emotions, we observe an absolute improvement in models' performance, but a relative decline toward the majority classifier benchmark. The best model, in this case, did not surpass the majority classifier in terms of accuracy, with the majority classifier achieving 0.67 ± 0.16 accuracy and 0.61 ± 0.16 F1-score, while SVM, the best-performing model, scored an accuracy of 0.64 ± 0.13 and an F1-score of 0.63 ± 0.12.

4.5 Feature Importances

Following the completion of model training, we analyzed the feature importances of the best-performing models. For RF this was calculated based on the Mean Decrease in Impurity, summing the impurity reduction each feature contributes across all trees; and for XGB, feature importances were calculated using the "weight" metric, which counts the number of times each feature is used to split the data across all trees. For SVM we did not calculate feature importances. In the completely personalised 10-fold experiments, feature importances varied significantly across different subjects and even between different runs within the same subject, specifically with RF, as the random state was not fixed. In contrast, feature importance was notably consistent in experiments where models were trained on data from multiple subjects, such as in the LOSO and the 10-fold within video, even with a variable random state of the RF model.

The most important features of best-performing models were those related to average pupil sizes, followed by fixation duration. These results partially align with those of Collins et al., who found features relating to pupil diameter and saccades statistically significant [3].

5 Conclusion

Our research explored emotion classification using eye-tracking data with classical ML models and hand-crafted features. The data was downsampled to a lower-than-standard frequency i.e., to 60 Hz, which was more realistic for consumer contact-free eye-tracker data. This made the problem harder, making it not directly comparable with other studies working on eSEEd, but valuable from a practical perspective.

Window segmentation significantly impacted model performance, with the best results constantly obtained using the largest window length. This suggests that longer observation periods capture more comprehensive information, making smaller windows less effective for emotion classification. We hypothesize that this does not transfer to realistic scenarios, as users might experience emotions in short bursts while being neutral for the majority of the time. In specifically designed cases where emotion is consistently induced for longer periods of time (like our dataset), this is more expected.

The LOSO validation strategy, which tests model generalization across different subjects, yielded poor results. The variability in performance across subjects indicates the challenge of capturing general relationships between eye features and emotions. While both 10-fold CV approaches showed an increase in performance, their generalizability is limited. Completely personalised 10-fold showed worse results than the marginally personalised one presumably because of the low number of videos per emotion within an individual subject.

Table 1: Best-performing models and their corresponding results along the results of the Majority Class Classifier for the same parameters. Window lengths are 10 s.

Settings	**Model Acc**	Model F1	Majority Class Acc	Majority Class F1
LOSO, RF	**0.28 ± 0.13**	0.28 ± 0.16	0.25 ± 0.25	0.15 ± 0.26
LOSO, SVM, negative emotions grouped	**0.59 ± 0.19**	0.46 ± 0.18	0.59 ± 0.19	0.46 ± 0.18
10-fold within video, RF	**0.60 ± 0.07**	0.60 ± 0.08	0.21 ± 0.01	0.07 ± 0.01
10-fold within video, XGB, negative emotions grouped	**0.76 ± 0.04**	0.73 ± 0.04	0.66 ± 0.02	0.52 ± 0.02
10-fold within subject, RF	**0.38 ± 0.20**	0.42 ± 0.19	0.33 ± 0.26	0.29 ± 0.26
10-fold within subject, SVM, negative emotions grouped	**0.64 ± 0.13**	0.63 ± 0.12	0.67 ± 0.16	0.61 ± 0.16

An important issue with the eSEEd data is that all participants watched the same 10 emotion-evoking videos in the exact same order. This uniformity raises concerns that, given the small number of videos (two intended[1] per emotion), the models might learn to associate features unrelated to emotions, such as video dynamics or illumination. We circumvented the problem with video dynamics by dropping the mean gaze coordinate features and not using them in our experiments.

Despite these challenges, our experiments offer valuable insights into the feasibility of emotion recognition from low-frequency eye-tracker data, providing a foundation for future work. We opted for classical models initially due to their explainability, lower computational complexity, and efficiency, which are in our opinion essential for understanding the data before transitioning to more complex deep learning models.

In future work, several enhancements could be explored to improve the robustness and accuracy of emotion classification models using eye-tracking data. One approach could involve analyzing distinct fixation areas as an additional feature, potentially offering deeper insights into visual attention patterns. Moreover, considering that each emotion is (in some cases) represented by two videos, a valuable experiment would be to train models on one video and test on the other. This could help assess the model's ability to generalize across different stimuli within the same emotional category.

Further analysis could focus on demographic factors by examining the LOSO results for potential correlations between model predictions and participant characteristics such as gender, age, and education. This might reveal underlying biases or trends that affect model performance. Additionally, rather than downsampling and removing rows with missing data, future work could explore retaining or imputing these rows.

Furthermore, exploring the training of neural networks on raw, non-downsampled data from multiple modalities is another promising direction, as other studies already observed promising results with such approaches. Moreover, we should address the issue of overlapping emotions which could involve developing a multiclass output model, reflecting a real-world scenario where multiple emotions can be present simultaneously. This approach could also help reduce the number of undefined labels, increasing the amount of useful data.

Acknowledgements

This work was supported by bilateral Weave project, funded by the Slovenian Agency of Research and Innovation (ARIS) under grant agreement N1-0319 and by the Swiss National Science Foundation (SNSF) under grant agreement 214991.

[1]The average percentage of the videos for which the participants had reported the target emotion (also known as the "hit rate") was 71.8% [13].

The authors acknowledge the use of OpenAI's ChatGPT for generating text suggestions during the preparation of this paper. All the generated content has been reviewed and edited by the authors to ensure accuracy and relevance to the research.

References

[1] Zeeshan Ahmad and Naimul Khan. 2022. A survey on physiological signal-based emotion recognition. *Bioengineering 2022*. https://www.mdpi.com/2306-5354/9/11/688.

[2] Aracena Claudio, Basterrech Sebastián, Snáel Václav, and Velásquez Juan. 2015. Neural networks for emotion recognition based on eye tracking data. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2632–2637. DOI: 10.1109/smc.2015.460.

[3] Mackenzie L. Collins and T. Claire Davies. 2023. Emotion differentiation through features of eye-tracking and pupil diameter for monitoring well-being. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. DOI: 10.1109/embc40787.2023.10340178.

[4] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. 2013. Differential entropy feature for EEG-based emotion classification. In *6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 81–84.

[5] Ralf Engbert, Lars Rothkegel, Daniel Backhaus, and Hans A. Trukenbrod. 2016. Evaluation of velocity-based saccade detection in the smi-etg 2w system. *Technical report, Allgemeine und Biologische Psychologie, Uni-versität Potsdam*.

[6] Atefeh Goshvarpour, Ataollah Abbasi, and Ateke Goshvarpour. 2017. An accurate emotion recognition system using ECG and GSR signals and matching pursuit method. *Biomed J. 2017*. DOI: 10.1016/j.bj.2017.11.001.

[7] Jiang-Jian Guo, Rong Zhou, Li-Ming Zhao, and Bao-Liang Lu. 2019. Multimodal emotion recognition from eye image, eye movement and EEG using deep neural networks. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 3071–3074. DOI: 10.1109/embc.2019.8856563.

[8] Robert Jenke, Angelika Peer, and Martin Buss. 2014. Feature extraction and selection for emotion recognition from EEG. *IEEE Transactions on Affective Computing*, 5, 3, 327–339. DOI: 10.1109/taffc.2014.2339834.

[9] Lim Jia Zheng, Mountstephens James, and Jason Teo. 2020. Emotion recognition using eye-tracking: taxonomy, review and current challenges. *Sensors 2020*. https://doi.org/10.3390/s20082384.

[10] Fjorda Kazazi. 2022. Detect saccades and saccade mean velocity in python from data collected in pupil labs eye tracker. Accessed: 25. 7. 2024. https://www.fjordakazazi.com/detect_saccades.

[11] Yousif Khaireddin and Zhuofa Chen. 2021. Facial emotion recognition: state of the art performance on FER2013. *arXiv preprint arXiv:2105.03588*. DOI: 10.48550/arXiv.2105.03588.

[12] Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu. 2015. Combining eye movements and EEG to enhance emotion recognition. In *Ijcai*. Vol. 15. Buenos Aires, 1170–1176.

[13] Vasileios Skaramagkas and Emmanouil Ktistakis. 2023. Esee-d: emotional state estimation based on eye-tracking dataset. *Brain Sciences*, 13, 4. DOI: 10.3390/brainsci13040589.

[14] Mohammad Soleymani, Maja Pantic, and Thierry Pun. 2012. Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing*, 3, 2, 211–223. DOI: 10.1109/t-affc.2011.37.

[15] Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski, and Remigiusz Jan Rak. 2020. Eye-tracking analysis for emotion recognition. *Computational Intelligence and Neuroscience*. https://onlinelibrary.wiley.com/doi/10.1155/2020/2909267.

[16] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. 2018. Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 28, 10, 3030–3043. DOI: 10.1109/tcsvt.2017.2719043.

[17] Wei-Long Zheng and Bao-Liang Lu. 2015. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7, 3, 162–175. DOI: 10.1109/TAMD.2015.2431497.

Indeks avtorjev / Author index

Andova Andrejaana 51
Anžur Zoja 31
Avdić Elma 15
Bengeri Katja 11
Bohanec Marko 59
Božak Tomi 83
Cigoj Primož 7
Cork Jordan 51
Đoković Lazar 19
Džeroski Sašo 67
Filipič Bogdan 51
Gams Matjaž 27
Gašparič Lea 67
Gjoreski Hristijan 35
Gjoreski Martin 35
Gradišek Anton 23, 79
Hafner Miha 59
Halbwachs Helena 23
Jelenc Matej 79
Jonnagaddala Jitenndra 79
Jordan Marko 71
Kalin Jan 27
Kizhevska Emilija 75
Kokalj Anton 67
Kolar Žiga 27
Konečnik Martin 27
Kramar Sebastjan 35, 71
Krstevska Ana 35
Kukar Matjaž 39
Kulauzović Bajko 27
Kuzman Taja 7
Lukan Junoš 11, 35
Luštrek Mitja 11, 31, 35, 71, 75, 83
Mehanović Dželila 15
Nedić Mila 63
Pavleska Tanja 7
Pejanovič Nosaka Tomo 27
Piciga Aleksander 39
Poljak Lukek Saša 47
Prestor Domen 27
Ratajec Mariša 23
Reščič Nina 71
Robnik-Šikonja Marko 19
Rupnik Urban 7
Sadikov Aleksander 43
Shulajkovska Miljana 79
Skobir Matjaž 27
Slapničar Gašper 31, 35, 83
Smerkol Maj 23
Šoln Kristjan 55
Susič David 27
Susič Rok 23
Trojer Sebastijan 31, 35
Tušar Tea 51, 63
Vladić Ervin 15

Založnik Marcel 55, 71
Zirkelbach Maj 43

Slovenska konferenca o umetni inteligenci

Slovenian Conference on Artificial Intelligence

Uredniki > Editors:

Mitja Luštrek, Matjaž Gams, Rok Piltaver