International Journal of Interactive Multimedia and Artificial Intelligence

June 2025, Vol. IX, Number 3
ISSN: 1989-1660

unir LA UNIVERSIDAD EN INTERNET

"Any sufficiently advanced technology is indistinguishable from magic."
Arthur C. Clarke

EDITORIAL TEAM

Editor’s Note

The International Journal of Interactive Multimedia and Artificial Intelligence (IJIMAI) publishes articles discussing the latest current topics in the research literature. The emergence of ChatGPT and other similar models based on deep learning are dramatically changing the way people understand and use artificial intelligence. Despite the significant advances made in these types of techniques, which have been enormous in recent years, new learning methods are still needed. Specifically, we require methods that allow us to handle data correctly in specific environments, as well as provide learning methods with the necessary explainability that allows us to understand how they are reasoning. The latter is essential for creating ethical learning methods that do not make unfair decisions based on biased information. It is also important to identify data that have, in some way, reflected the reprehensible attitudes and reasoning that we as fallible human beings sometimes have. In short, artificial intelligence should reflect, if possible, the best of us rather than the worst. With this goal in mind, it is common to see in this issue of the journal an abundance of articles proposing new learning methods, many of which are based on Deep Learning and Data Mining. There are also articles on large language models, which are extremely important in the current artificial intelligence landscape. Of course, there are also articles on optimization methods and quantum computers, which are also of great importance in the field of artificial intelligence. Although generative artificial intelligence models are perhaps the ones that have people most intrigued, this is not the only current application of artificial intelligence. We are seeing how renewable energies, in particular those that come from the sun and wind, are playing an increasingly important role in global energy generation. As seen in recent events, such as the general blackout in Spain, the electricity system needs new methods that allow adequate regulation to prevent all kinds of possible failures. In this issue, two articles present new applications of artificial intelligence methods to renewable energy generation systems. Also noteworthy within this issue is the application of artificial intelligence in the field of teaching, where the aim is to provide a better learning experience for students and teachers.

This issue of the journal begins by reviewing the advances and challenges in AI-generated text detection. As mentioned, the rapid development of AI in recent times has raised many ethical issues. One such case involves fraud and the use of AI-generated texts as if they were one's own. Solving this problem and developing new effective detectors to identify such cases are vital for the correct and ethical use of artificial intelligence.

Next, we present an article demonstrating the great advantages that the use of new large language models can provide for the augmented reality field. Specifically, the authors proposed a new method of assistance for the understanding and easy documentation of these environments by expert users. Traditionally, the field of augmented reality requires experts to enter information using structured formats, which, for them, are rather tiresome to use. The use of large language models makes it possible to simplify this task and add information to augmented reality environments in a simpler and more convenient way.

The authors of the following article studied the effectiveness of training recommender systems based on Deep Learning using synthetic datasets created from real datasets. Specifically, the authors test the Generative Adversarial Networks for Recommender Systems (GANRS) method on 3 synthetic datasets created from 3 different real datasets. Among other experiments, the authors compare the effectiveness of GANRS against 6 other Deep Learning methods considered state-of-the-art in the field. The results demonstrate how the proposed GANRS method generates consistent results for the datasets used.

The fourth article in this issue describes data management in sensitive environments. Specifically, it deals with the use of artificial intelligence patient diagnosis in emergencies. Although it is, of course, always best to have a doctor examine the patient, there are many situations in which the patient cannot wait for the necessary doctor to be available. It is in these cases that artificial intelligence can provide critical aid. Therefore, an explainable artificial intelligence-based disease diagnosis and blockchain-based decision-making system is proposed to address these challenges and improve patient care. The authors of this article propose solutions for handling unstructured clinical information, which provides high-quality information to the system and allows reliable responses to be generated. In addition, the use of Blockchain technologies prevents erroneous decisions, and the solution must be verified by at least 50% of the experts. For decision-making, the authors have chosen a recommendation system based on ant colony optimization.

Continuing with the processing and analysis of medical and personal data, the authors of the following article propose the use of artificial intelligence for humanitarian purposes. Specifically, the authors used machine learning techniques to detect patterns among homeless people suffering from drug addiction. To do so, they used real data from the National Administrative Department of Statistics (DANE) of Colombia. Specifically, 19375 records and 25 columns. The results obtained in this article will allow municipal administrations to make decisions that will help improve the situation of these people.

The next two articles in this issue use optimized Long Short-Term Memory (LSTM) neural networks for different purposes. The first one addresses the problem of estimating the amount of energy generated in a photovoltaic power plant. For this purpose, this study focused on a photovoltaic installation with 296 panels located in the northwest of Spain. Synthetic data were used to train the neural network for the estimation. In the second article, the authors propose a method for estimating wind power using optimized LSTM neural networks. Estimating the wind power is essential because it helps to estimate the energy generated in a wind power plant. In addition to the neural network, data pre-processing was applied. Concretely, two techniques such as removal of missing values and inputting missing values using Random Forest Regressor (RFR), are used.

One of the most important problems when applying data mining is handling missing values. The authors of the next article propose a methodology for evaluating different estimation techniques for missing values. In this evaluation, they consider the use of quality metrics derived from data mining processes. To do so, they compared the effectiveness of the data mining methods when applied to complete datasets and when applied to the same datasets but with missing values. Specifically, the authors apply this test on 63 different datasets using the median, K-Nearest Neighbors (KNN), K-means, and Hot-Deck imputation methods.

Next, in this issue, we include two articles in which artificial intelligence was applied to solve vehicle traffic-related problems. In the first article, the authors focus on traffic optimization. With the increase in the world’s population, the ease with which people can access a motor vehicle, and the growing population of cities, it is necessary to develop new methods to optimize traffic. These methods must allow everyone to reach their destination in the shortest possible

DOI: 10.9781/ijimai.2025.05.002

time and, at the same time, avoid possible traffic jams and accidents. With the aim of optimizing traffic, the authors present a method that, by using evolutionary algorithms and waiting time prediction, attempts to optimize traffic. The proposed method combines the use of two different techniques. On the one hand, the waiting time of vehicles is estimated using a set of techniques, and on the other hand, using the calculated information, evolutionary algorithms are applied to generate the final optimization. The combination of these techniques allows its use in real-time. The proposed method was successfully tested in real situations.

The following article on this issue focuses on the care of children who are victims of road accidents. Road accidents are becoming a problem that requires immediate decisions and care for the injured. In order to improve the health services treatment procedures, the authors of this article propose a new clinical decision-making method based on case-based reasoning and data mining to streamline and improve the care of children injured in road accidents. The aim of the article is to develop an efficient predictive model to determine whether or not a child victim of a traffic accident should be admitted to a pediatric intensive care unit. The proposed method preprocesses data using the KNN method. For its evaluation, real data elaborated by the authors and validated using statistical analysis techniques were used. The results were positive with a hit rate of 91.66%.

The next article focuses on one of the main ethical problems that new artificial intelligence methods generate. Specifically, we refer to the personal data treatment. All types of Internet websites and entities of all kinds usually collect data from their users. Nevertheless, it is often very difficult for an average user to know exactly what they are approving of by providing personal data. All the information that entities provide to users usually consists of long documents that users generally do not have time to read or do not understand. In this article, the authors present a system of icons that aims to make it easier for users to understand what they are accessing when providing personal data. Specifically, this system is designed for an academic environment to retrieve information from a set of students that apply to online courses. In addition, the authors designed a survey system to determine whether students understood the proposed icon system or not. This allows the icon system to be subsequently refined and adapted to the users, in this case, the students of the courses.

Following on from the previous article on the application of artificial intelligence in the field of teaching, the next article in the journal presents a study in which the interactions of a series of students with a series of class exercises are analyzed. Using machine learning techniques, it has been analyzed how a student's grade can vary depending on the time spent on the exercises and the number of attempts to solve them. The results demonstrate that for exercises with an average number of attempts of 2, the model converged in 200 iterations. It was also observed that the probability that the student gets the exercise right randomly is very low.

The next article of this issue discusses image protection. In it, the authors propose a new method for creating reversible watermarks using the Modified Quadratic Difference Expansion and Hybrid Optimization Technique. Thanks to this method, it is possible to protect an image with a watermark; thus, the image can be removed if there is sufficient authority. The proposed method proceeds as follows. First, fractal encryption is applied to watermarks using Tromino's L-shaped theorem to improve security. Next, Cuckoo Search-Gray Wolf Optimization (CSGWO) is applied to the cover image to optimize block allocation in order to insert the watermark image. The proposed method achieved an average Peak Signal-to-Noise Ratio (PSNR) of 60 dB.

The following article on the issue examined the reliability of IBM's public quantum computers. Specifically, the authors monitor the reliability of IBM's public-access QCs network daily. The study machines have different qubit associations. For the testing, the authors employed an ad hoc computationally demanding quaternary search algorithm that they executed every 24 hours for 100 days. The main reason for this is to limit the operational capacity to its limits. The authors then performed a comparative analysis considering similarities and the total number of executions. Moreover, the authors applied 50 days of improvement filtering in order to mitigate the noise in the system. The Yorktown 5-qubit computer achieved noise filtering of up to 33% in one day, that is, a 90% confidence level was reached in the expected results. For long-term tests, the authors concluded that improvement is needed.

In the penultimate article of the issue, authors present a muti-session evaluation of a haptic device in order to compare its performance in normal and critical conditions. The idea is to test a device that will be used by astronauts on future missions to the Moon and Mars. For the test, 8 different factors have been taken into account, 6 groups of astronaut pairs were created and 4 test sessions were conducted. In addition, the experiment was recreated under stressful conditions, including a session in which a critical condition simulated in an extra-vehicular situation.

The last article reports on a comprehensive comparative analysis of four machine learning algorithms for customer segmentation in the retail sector. Specifically, the authors used two datasets, a large-scale Turkish market sales dataset and a focused marketing campaign dataset. K-means showed a robust performance, offering a balance between interpretability and statistical validity. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) showed strengths in identifying non-spherical clusters and handling outliers, while Gaussian Mixture Models (GMM) and Self-Organized Maps (SOM) provided more granular segmentation but increased complexity. By introducing a methodological framework for the evaluation of customer segmentation techniques, this study enhances current practices in retail analytics.

As closure, I would like to thank the authors and reviewers for their hard work. Without them, publication of this issue would definitely not be possible. I would also like to thank all the readers of the journal for their continued interest.

Dr. Juan Antonio Morente Molinera
Associate Editor

TABLE OF CONTENTS

OPEN ACCESS JOURNAL

ISSN: 1989-1660

The International Journal of Interactive Multimedia and Artificial Intelligence is covered in Clarivate Analytics services and products. Specifically, this publication is indexed and abstracted in: *Science Citation Index Expanded, Journal Citation Reports/ Science Edition, Current Contents®/Engineering Computing and Technology.*

Distinguishing Human From Machine: A Review of Advances and Challenges in AI-Generated Text Detection

Serena Fariello, Giuseppe Fenza, Flavia Forte, Mariacristina Gallo, Martina Marotta *

Department of Management and Innovation Systems, University of Salerno, 84084 Fisciano (SA) (Italy)

* Corresponding author: s.fariello4@studenti.unisa.it (S. Fariello), gfenza@unisa.it (G. Fenza), f.forte27@studenti.unisa.it (F. Forte), m.marotta37@studenti.unisa.it (M. Marotta).

Received 22 July 2024 | Accepted 28 October 2024 | Published 18 December 2024

Abstract

The rise of Large Language Models (LLMs) has dramatically altered the generation and spreading of textual content. This advancement offers benefits in various domains, including medicine, education, law, coding, and journalism, but also has negative implications, mainly related to ethical concerns. Preventing measures to mitigate negative implications pass through solutions that distinguish machine-generated text from human-written text. This study aims to provide a comprehensive review of existing literature for detecting LLM-generated texts. Emerging techniques are categorized into five categories: watermarking, feature-based, neural-based, hybrid, and human-aided methods. For each introduced category, strengths and limitations are discussed, providing insights into their effectiveness and potential for future improvements. Moreover, available datasets and tools are introduced. Results demonstrate that, despite the good delimited performance, the multitude of languages to recognize, hybrid texts, the continuous improvement of algorithms for text generation and the lack of regulation require additional efforts for efficient detection.

Keywords

Generated-Text Detection, AI-Detection, Large Language Models (LLMs), Literature Review, Survey.

DOI: 10.9781/ijimai.2024.12.002

I. Introduction

With the increase of computer power and the availability of extensive datasets, Artificial intelligence evolve rapidly. In the area of Natural Language Processing (NLP), the introduction and diffusion of Large Language Models (LLMs) have transformed existing approaches due to their ability to achieve significant performance in different NLP tasks. In the early days, this included simply automated responses in customer service, conversation summaries like automated call transcriptions, news articles, and so on. As a result of technological evolution, machine-generated text has become more and more sophisticated and human-like [1]: modern systems use advanced algorithms and analyze vast amounts of data to produce natural and coherent text [2]. They are utilized in varied contexts and for different purposes: writing articles, providing customer support, and even creating educational content, as described in the following. This advancement led to the development of Large Language Models (LLMs), which have significantly changed the way in which people generate and interact with machine-produced text. LLMs reveal a significant capacity to generate text that matches human writing. This capacity makes distinguishing LLM-generated text from human-written text hard. However, the machine-generated text could impact ethical issues such as exacerbating biases and stereotypes in training data or producing false or misleading content. Known issues related to machine-generated content rely on manipulating public opinion, spreading fake news, and plagiarism. So, despite the huge potential, it is important to use LLMs conscientiously to avoid cheating, dishonesty and low-quality responses [3] [4]. Preventing measures aiming to mitigate future implications of LLMs diffusion are necessary and pass through valid solutions distinguishing machine-generated text from human-written text. The present study intends to collect and analyze the most recent approaches in terms of detection and identification of generated text content. In particular, the research work aims to answer the following questions:

RQ1 What are the most recent methods for detecting LLM-generated texts and their main limitations?

- Various detection methods are reviewed, including watermarking, feature-based, neural-based, hybrid, and human-aided approaches.
- Each method's strengths and potential areas for improvement are highlighted.

Please cite this article as:
S. Fariello, G. Fenza, F. Forte, M. Gallo, M. Marotta. Distinguishing Human from Machine: A Review of Advances and Challenges in AI-Generated Text Detection, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 6-18, 2025, http://dx.doi.org/10.9781/ijimai.2024.12.002

RQ2 What datasets are used for training detection models?

- The study examines the datasets utilized for training detection models.
- The advantages and limitations of these datasets in accurately identifying machine-generated texts are discussed.

RQ3 Are there state-of-the-art tools capable of addressing recent advancements in text generation?

- The study evaluates the effectiveness of current detection tools.
- Emphasizes the need for continuous development to keep pace with advancements in LLM capabilities.

The rest of the manuscript is structured as follows: Section II overviews LLMs' functioning, their applications, motivations guiding this research work, and a focus on the targeting task: *machine-generated text detection*. Section III examines the existing literature review in the machine-generated text detection task, and Section IV outlines the research methodology. Section V delves into the detection methods, categorized into watermarking, feature-based, neural-based, hybrid and human-aided approaches. Section VI describes the characteristics of existing datasets and discusses their limitations, while some useful tools are examined in Section VII. Finally, Section VIII discusses the problems and limitations of examined detection methods, and Section IX concludes the manuscript.

II. Context and Background

LLM-generated text is the latest and most advanced form of machine-generated text. These models, such as GPT-4 and BERT, use deep learning to produce texts extremely close to those written by humans. The models are trained on massive datasets, including a huge variety of human language examples, allowing them to understand and mimic complex patterns, syntax, and meanings [5].

This section introduces LLMs, their application, and the motivations that guide this research work. Moreover, the objective of the considered literature is detailed.

A. LLM Fundamentals

Large Language Models are based on an advanced neural network, especially the Transformer architecture introduced by Vaswani et al. [6]. It is based on the following equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Each word in the sentence is converted into a numerical vector. Q (query) is the word vector, K (key) are the vectors for every word in the sentence, and V (value) are the value vectors. Softmax is the function that converts scores into a probability distribution. This formula allows the model to weigh the importance of each word within a sequence of other words.

Generation of text (contextually relevant and coherent) by LLMs like Xlnet depends from their autoregressive nature. Aurogressive models predicts future behaviour based on past behavior data. In the case of text, the subsequent word is predicted based on the previous ones [7].

B. Motivations

Due to their performance, LLMs are utilized in varied contexts and for different purposes: writing articles, providing customer support, and even creating educational content. The generated text is so impressively good that it is difficult to tell if it was written by a person or a machine. Nevertheless, LLMs could produce inaccurate information. This creates a challenge in identifying AI-generated content, which has led to the development of advanced detection techniques. The following are examples of LLMs' applications and their risks:

- **Education**. LLMs can provide personalized and interactive learning experiences for students and may help teachers reduce their workload in order to focus on research [8]. According to Jeon and Seongyong [9], LLMs such as ChatGPT may help teachers by assuming supporting roles like interlocutor, content provider, teaching assistant and evaluator. However, according to specific subjects, LLMs' performance and responses may have different accuracy grades. For instance, in Geometry [10], LLMs sometimes cannot provide accurate and reliable answers due to a lack of critical and logical thinking, leading to the necessity of human integration.
- **Medicine**. LLMs are starting to be used in the healthcare sector to enhance the well-being of both patients and doctors. ChatGPT and Med-palm 2, for example, have exhibited encouraging outcomes in medical assessments and addressing patient inquiries, even if they are still imperfect and have shown a lack of recency, accuracy and coherence. Therefore, at present, they cannot be deemed as a true replacement for medical professionals but rather as a supplementary tool in clinical, educational, or research environments [11]. They faced challenges in understanding cause-and-effect relationships between medical conditions and lacked sufficient medical knowledge to fully comprehend complex interactions [12].
- **Coding**. In the realm of software development, where creating applications involves writing code in various programming languages, the rapid progress of LLMs is proving beneficial. Feng et al. [13], in their research, have discovered that ChatGPT has been employed across many different languages - with Python and JavaScript emerging as the most widely utilized - for different coding tasks like debugging and testing. However, unlike common writing assignments, programming requires precise conformity to syntax and rules and great attention to possible vulnerabilities, making it notably challenging for generative models to produce top-notch and high-security code [13],[14].
- **Law**. LLMs are transforming how legal professionals work, enhancing their efficiency and accuracy in daily tasks such as legal research, contract drafting [15], empirical analysis (LLMs can be used to examine large volumes of legal texts, identifying trends and arguments) [16], assistance in contract negotiation and creation of legal contents [17]. However, there remains a significant risk of relying on inaccurate, outdated, or unsourced legal information [18].
- **News Generation**. Nowadays, journalists use LLMs to fabricate news to maximize the spread of content and take advantage of social networks. Nevertheless, the risk is a compromised authenticity of news as well as biased content [19].

As outlined, machine-generated content, on one side, can improve and facilitate the work; on the other side, it can undermine academic and journalistic integrity, intellectual property [20], transparency, and ethics [21]. Knowing the state-of-art in terms of solutions to recognize the nature of content could help in developing more suitable solutions, regulating the use of genAI [22] and, finally, improving generative models themselves.

C. Machine-Generated Text Detection Task

This literature review intends to collect and discuss the state of the art in terms of approaches for detecting machine-generated text. The machine-generated text detection task consists of automatically detecting content generated by LLMs. It can be solved as a binary classification problem or by setting a threshold. From a mathematical perspective, it can be formalized as a binary classification problem, seeking to determine if a given text is generated by an LLM or by a human writer [5].

Given a text t and a Detector $D(t)$, the equation is the following:

$$D(t) = \begin{cases} 1 & \text{if } t \text{ is machine} - \text{generated}, \\ 0 & \text{if } t \text{ is human} - \text{written} \end{cases}$$

Setting a threshold can contribute to another way to define the detection task. Given an input text, the text detector outputs a score. A score higher than the threshold indicates a machine-generated text [23].

III. Related Works

The detection of generated text is a hot research field to explore. In fact, several works in the literature have reviewed the main techniques used to perform this task. One of the early technique reviews goes back to 2016 [24]. From then on, the wide ever-increasing use of LLMs definitely complicated generated text detection. As a consequence, the research started to be more attentive to the text generated by these models, and scientists began investigating and publishing new machine-generated text detection methods [25]. With ChatGPT's rise, there was a further increase in LLM-generated text reviews, such as the approach proposed by Dhaini et al. [26]. This research line has become the main one as it has been supported by many different works, which have inspired this work as well. In detail, a further step has been made by a work that introduces a first kind of categorization by dividing feature-based and neural-language model approaches [27]; another work [28] divides the task into black-box and white-box detection, introducing a novel template followed by successive works; another method [29], following the black-box and white-box structure idea, added three categories of detection methods: training-based, zero-shot-based and watermarking methods. Moreover, interesting research [23] highlighted the weaknesses of existing text detection techniques (e.g., text paraphrasing). At the same time, the review work by Uchendu et al. [30] has introduced the hybrid methods category for the first time. The most exhaustive study, considering the state-of-the-art methods to date, is the one proposed by Wu et al. [5] that covers many method categories. It presents useful sections for the ones dealing with the phenomenon of generated text detection, providing details regarding the most popular datasets and benchmarks useful for this task and underlining the research limitations in generated text detectors.

The proposed survey begins by exploring various risks associated with multiple application domains of machine-generated text, underscoring the urgent need for robust detection methods. Compared to previous surveys, it is more up-to-date and provides a thorough analysis of the strengths and weaknesses of current approaches. In addition, the survey offers an in-depth discussion of the datasets used to train learning models, highlighting current limitations in terms of data quality and diversity, and summarizes the performance evaluation of state-of-the-art tools. Finally, emerging challenges, such as issues related to the adopted languages and the lack of regulatory frameworks, are discussed.

IV. Research Methodology

The papers guiding this study are harvested from relevant search engines like Scopus, DBLP, and Scholar by exploiting specific queries: *LLM-generated text detection*, *machine-generated text detection*, and *authorship attribution*. Moreover, Scimago has been adopted to filter more relevant journals, while the International CORE Conference Rankings (ICORE) for conferences. The authors' H-index was considered in the case of very recent preprint versions. During the collection, journals with an h-index higher than 10 and conferences with a performance class that ranges from A to B were considered. Concerning preprints, the works created by authors with an h-index higher than 10 or those with a number of author citations higher than 35 were selected. Fig. 1 shows the distribution, by year, of the last six years of literature on the generated text detection task, highlighting a peak after the introduction of GPT. Results are summarized in Table I.

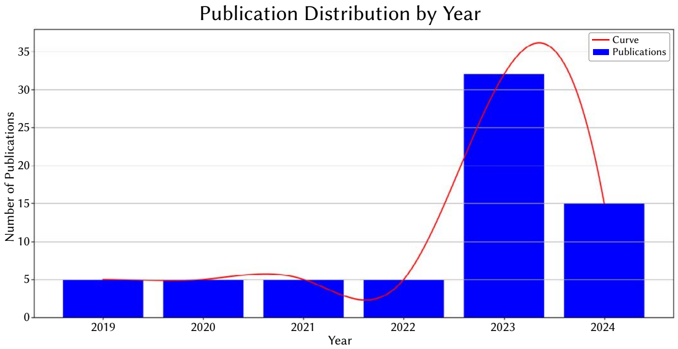


Fig. 1. The distribution, by year, of the last six years of literature on the generated text detection task.

V. Generated Text Detection Methods

This section analyzes the detection methods emerging from the analysis in detail and arranges them into five categories: watermarking, feature-based, neural-based, hybrid and human-aided approaches.

A. Watermarking

Watermarks are embedded signals in the generated text that are invisible to humans but can be detected involving the use of algorithms. Text watermarking implements patterns into the generated text to tell the difference between large language models (LLMs) generated text and a human-generated text [36],[68]. Watermarking must be effective (the coherence of the generated text must be preserved), invisible (it should smoothly blend into the text), robust (it requires being difficult to eliminate [28] resisting to corruption or attacks [42]).

1. Data-Driven Watermarking

The aim of data-driven methodologies is to assess the property of data. Using patterns or tags within the training datasets these methods can check if the data is copied or used for malicious ends. Adding a few samples with hidden watermarks (backdoor insertion) to the training data can help language model creators detect if their models are being used by bots on platforms like Twitter to spread fake news. Thus, the model learns a secret function set by the creator. The watermark is very robust even in case of a model being fine-tuned for other specific tasks [31].

Succeeding studies have identified weaknesses in this technology, revealing that it can be easily manipulated.

Lucas et al. [32] found that inserting triggers made from uncommon markers makes them difficult to detect.

Triggers constructed with ordinary words are less effective for watermarking because the presence of common word combinations in natural text poses a risk of false positives, text incorrectly detected as generated by a language model when it was actually authored by a human. Additionally, watermarks based on common words are way easier to detect. Research by Tang et al. [33] demonstrates that incorporating just 1% of watermarked samples improves traceability in

TABLE I. Summary of Evaluated Approaches

Category	Sub-categories	Overview	Advantages	Limitations
Watermarking	Data-Driven [31][32] [33]	Allows checking whether a given text has been generated by a model that uses a watermark.	Effective against attempts to remove or modify it. Compatible with any LLM. Minimal to no effect on the quality of a generated text.	Foolable through paraphrase. Needs the willingness to apply.
	Model-Driven [34][35] [36] [37]			
	Post-Processing [38] [39] [40] [41][42] [43]			
	Neural-based [44]			
Feature-based	Stylistic Feature and Stylometry [45] [46] [47]	Leverages measurable and evident differences between human and generated texts in terms of syntax, grammar, and other linguistic particularities.	Flexible and less computationally intensive. Makes the decision-making process transparent and comprehensible.	Susceptibility to perturbations (e.g., word substitutions). Difficulty in transferring features across architectures.
	Frequency Features [48]			
	Statistical Metrics [49] [50] [51]			
Neural-based	Feature-based [30][52] [53] [54]	Exploits the architecture of deep neural networks.	Effective in capturing the complex linguistic nuances present in texts. Robust against small mutations in text. Efficient in different application scenarios.	Strictly related to domains and languages of training datasets and to the adopted model for generating text.
	Pre-training [54][55]			
	Fine-tuning Classifier [56] [55]			
	Zero-shot [57] [56][58]			
Human-aided [59] [60] [61] [62] [63] [64]		Combines features-based or neural techniques with human analysis and review.	Enhanced accuracy thanks to human review.	Strictly related to human skills.
Hybrid [65] [66] [67]		Combines multiple methodologies.	Hard to obfuscate the style and deceive detection systems.	High complexity and computational costs.

datasets, facilitating better management and safeguarding of language models. It should be noted that data-driven approaches primarily aim to safeguard dataset copyrights, thus typically offering limited payload capacity and applicability. Furthermore, implementing these methods in detecting text generated by LLMs demands substantial resources, such as embedding watermarks across extensive datasets and retraining the models.

2. Model-Driven Watermarking

This kind of method integrates watermark signals directly into Large Language Models. They do so intervening on the logits distribution or on the token sampling.

Logits-Based Methods

The watermark of Kirchenbauer et al. [34] comes in the decoding step. So before choosing the next word, watermarking randomly excludes a portion of the possible words (blacklisted). Limiting the model's choice to the remaining options (whitelisted). The seed for the random number generator that chooses which words are blacklisted is the last word of the input. In this way, the blacklist can be reconstructed at any time. This procedure is applied at each generation of the next token. To detect generated text by a language model, one needs to detect the watermark counting the blacklisted words in the generated text. Obtaining the blacklist means knowing the random number generator used to choose the blacklist words and the seed. The watermarked language model would not use blacklisted words because it can not, but humans would definitely use blacklisted terms. So, a text using only whitelist words is highly likely to be AI-generated, and even a short text can be classified with relatively high certainty. Recent research conducted by Kirchenbauer et al. [35] demonstrates that watermarking remains effective even when watermarked text is manually rewritten, paraphrased by non-watermarked LLMs, or integrated into longer handwritten documents.

However, to generate and detect watermarks it is needed a secret key poses potential security vulnerabilities. In response to this matter, a research [36] introduced the first private watermarking algorithm. This method employs separate neural networks for watermark generation and watermark detection. In this way, two different keys can be used. Additionally, both networks share a section of the parameters, enhancing the detection network's efficiency and accuracy. Existing watermarking methods for LLMs only contain one bit of information (whether it is generated from an LLM or not) and cannot flexibly give information such as model version, generation time, user ID, etc. In this sense, Wang et al. [68] conducted the first study on the topic of Codable Text Watermarking for LLMs (CTWL) that allows text watermarks to carry more customizable information including which model generated the text and when.

Token Sampling-Based Methods

Token sampling on language models represents the process of selecting subsequent words (tokens) following a probability distribution. Token sampling entails randomness so that the resulting text becomes unpredictable. Methods utilizing token sampling for watermarking use random seeds or specific patterns to guide the token sampling mechanism. The method proposed by Kuditipudi et al. [69] used a secret key, which is a set of random numbers, to control token sampling. This token sampling operation is incorporated into the language model so that the output text contains an embedded watermark. To detect watermarks, the confidential key is used to line up the text with the arbitrary numbers. This hidden number makes it possible to recognize and recover the watermarks from the watermarked text. Paraphrasing would be difficult in this technique. Another recent work is SemStamp [37], which involves the use of Locality-Sensitive Hashing (LSH) to watermark sentences generated by the language model. Locality-Sensitive Hashing is a method that maps similar points in semantic space to adjacent positions in hash space. It subdivides the semantic space into two regions: one with watermarks and another without watermarks. This facilitates the identification of watermarked sentences during the detection phase. From the experimental results, SemStamp is more robust when it comes to the common type of paraphrasing attempts that involve two adjacent words than the other existing methods and is more effective in maintaining the quality of text generation.

3. Post-Processing Watermarking

Post-processing watermarking is the practice of adding a watermark by modifying the generated text by a LLM.

Character-Level Methods

In the past, watermarking was done by inserting or substituting unique Unicode characters in a piece of text. With these techniques, the characters convey encoding information but they are invisible to the human eye.

Lip Yee Por et al. [38] proposed UniSpaCh, a method for hiding data in Microsoft Word documents using Unicode characters, which enhances embedding efficiency and resists attacks while preserving the document's original appearance.

Word-Level Methods

Yang et al. [41] proposed a natural language watermarking scheme based on context-aware lexical substitution. They employ BERT [70] to suggest lexical substitution candidates by inferring the semantic relatedness between the candidates and the original sentence. A watermark insertion model [40] detects alterations in the text even in the presence of paraphrased content. The process of watermarking insertion is based on a methodology that selects and replaces words with synonyms to embed watermarks in sentences while preserving grammatical integrity. They also use BERT for watermarking detection because it possesses the capability to recognize sentence modifications and distinguish between marked and unmarked sentences. In another work [42], features like proper nouns and words' grammatical dependency, that are semantically or syntactically fundamental components of the text and, thus, invariant to minor modifications in texts, are identified and used as anchor points to pinpoint the position of watermarks. It is a multi-bit watermarking framework able to embed adequate bits of information and extract the watermarks in a robust manner despite possible corruption, such as copy-paste attacks, substitution attacks, paraphrasing attacks, etc.

Yang et al. [43] present a method that uses a binary encoding function. This function associates binary codes to words, in an arbitrary manner. For example, a word like "happy" could be replaced with "joyful" if it represents a "1" in the binary code, while "sad" might remain unchanged if it represents a "0".

Neural-Based Approach

An Adversarial Watermark Transformer (AWT), an innovative system that automates the entire process of embedding watermarks into texts, is proposed [44]. It utilizes the Transformer to learn how to replace specific words with others that carry a secret binary message. With this approach, the algorithm handles everything, from selecting words to embedding the watermarks, without the need for manual intervention. Additionally, AWT leverages adversarial techniques, meaning it trains itself to be resistant to attempts at watermark detection and removal.

Three watermark networks are taken into consideration in neural-based approaches: an encoder, a decoder and a discriminator.

The encoder generates a modified text that incorporates the watermark, which can be a sequence of binary bits. The modifications in the text must be minimal to preserve the readability and naturalness of the text. The decoder extracts the watermark from the modified text produced by the encoder network. The discriminator distinguishes between the original text and the watermarked modified text. Its aim is to prevent the encoder from significantly altering the text. During training, the discriminator tries to identify which text has been modified by the encoder while it attempts to make the modified text as similar as possible to the original text to fool the discriminator.

The performances are considered satisfactory if the encoder successfully embeds the watermark into the text in a way that makes it difficult to detect, the decoder is capable of accurately retrieving the message and the discriminator canâ€™t notice the difference an authentic text and a watermarked one.

4. Advantages of Watermarking

Watermarking is an adequate choice for many reasons. A watermark remains effective even when attempts are made to remove or modify it [34]. Additionally, it is compatible with any large language model and has minimal to no effect on the quality of the generated text.

5. Limitations of Watermarking

Despite their applicability, there are multiple ways to fool watermarking algorithms. By knowing the blacklist, the tokens in it can be used within the text. But brute-forcing their way to the blacklist means that the attacker queries the API a lot of times with the same input, in which case, the API provider can monitor and detect this malicious activity. Another way to attack watermarking is by doing word substitutions: the rewritten text will not be detected by watermarking. The attacker could also use a non-watermarked model to paraphrase the output of a watermarked model. Making minor adjustments, such as inserting spaces, emojis, or misspellings, can impact watermark detection. The main disadvantage of watermarking is that it can only be implemented when individuals and organizations are willing to apply it to their language models. In addition, existing tools are applicable to language models that do not implement watermarking. Future strict regulations about it could help implement this technique.

B. Feature-Based Methods

Feature-based methods leverage the fact that there are measurable differences between human and AI-generated texts in terms of syntax, grammar, and other linguistic particularities.

Munoz-Ortiz et al. [71] made a quantitative analysis comparing human-written English news text with output from LLMs of the LLaMa family. Their research leads to important discoveries about:

- **Sentence Length Distribution**: Human texts exhibit more scattered sentence length distributions compared to LLM-generated texts.
- **Dependency and Constituent Types**: Human texts show a distinct use of dependency and constituent types.
- **Emotions**: Human texts display more aggressive emotions (e.g., fear and disgust) than LLM-generated texts.
- **Language Characteristics**: LLM outputs use more numbers, symbols, and auxiliaries than human texts. Additionally, LLMs employ more pronouns.
- **Sexist Bias**: The sexist bias prevalent in human texts is also expressed by LLMs.

Following the listed aspects, the current section shows the evolution of feature-based methods in distinguishing between human and AI-generated content, categorizing features into three main types: Stylistic Features, Frequency Features, and Statistical Metrics.

1. Stylistic Features and Stylometry

Stylometry is the quantitative analysis of literary style. It involves examining various linguistic features, such as specific vocabulary and verbs, syntax, sentence structure and length, fluency and consistency, to identify patterns and similarities/differences between texts or authors. It is commonly applied in fields such as forensic linguistics, literary studies, computational linguistics and authorship attribution.

Stylometry demonstrated its effectiveness in spotting fake news written by humans, but it has not the same effects in spotting fake news generated by machines [46]. Thus, to achieve more reliable results, it may be useful to integrate Stylometry with other methodologies. For instance, Abiodun Modupe et al. [45] have proposed a method called RDDN that uses a neural network to extract lexical stylometric features. These stylometric features are fed into a bidirectional encoder to generate a vector representation of syntactic features, and the vector is then used by a bidirectional decoder to learn the writing style of an author.

In a recent study, Kumarage et al. [47] presented an algorithm using stylometric signals to measure stylistic changes in human and AI tweets to detect AI-generated tweets.

Their experiments succeeded in showing that stylometric features (specifically the ones related to phraseology, punctuation, and linguistic diversity) effectively enhance AI-generated text detection.

Other examples of mixed methodologies based on stylistic features will be presented and discussed in Section E.

2. Frequency Features

Frequency features refer to the repetition of specific terms, the distribution of word sequences, the number of punctuation marks, and the frequency of grammatical errors [27]. In their work, Frohling et al. [48] developed a feature-based classifier that leverages various features, including those related to the concept of repetitiveness. This can be measured by counting the number of stop-words, unique words, and words from ”top-lists” within a text. They specifically looked at the overlap of n-grams for words (lexical repetition) and part-of-speech tags (syntactic repetition) in consecutive sentences, under the assumption that human text tends to be less repetitive than generated text in both sentence structure and word choice.

3. Statistical Metrics

Metrics such as perplexity and entropy are crucial in evaluating the predictability and variability of a text. This section exposes some examples of how perplexity, burstiness, entropy and density can be used in a generated-text detection task.

Perplexity is a metric that measures how well a probability model, such as an LLM, predicts a sample.

Specifically, it assesses the model’s uncertainty in predicting the following word to choose, based on the preceding words, to continue a phrase.

Language models token sampling depends on common patterns in the training data. Therefore, LLM-generated text is characterized by low perplexity. In contrast, humans express themselves using non-identical styles, exhibiting higher perplexity values [49].

Burstiness measures the sentence complexity. Humans vary their sentences a lot when judging by the length and the number of rare words they use. So, burstiness has something to do with the fact that, for example, rare words usually do not occur very often in writing, but when they do, they start to happen a lot for a sentence or two, then not anymore. Language models are more constant in the way they write out their sentences. So going on sentence by sentence, one can plot the complexity of each sentence. For humans, these values will vary a lot, while for models, the value will be quite similar for all sentences. Then, a bumpy burstiness graph will likely belong to a human text, while a more constant graph will belong to an AI-generated text. GPTZero[1] is an example of a tool applying Perplexity and Burstiness to detect AI-generated text content.

While perplexity focuses on the model’s predictive accuracy for specific word sequences, ***Entropy*** quantifies the overall uncertainty and randomness in word distribution across a text. In the past, researchers have shown that human-written texts generally exhibit higher entropy due to their varied word choices, whereas AI-generated texts often demonstrate lower entropy as they tend to follow more structured conventions [59]. However, as models like GPT-3 and GPT-4 advanced, they became capable of generating more diverse and contextually rich text, closely mimicking human variability and resulting in higher entropy. Recent findings by Mitchell et al. [50] support this new perspective. They observe that entropy correlates positively with the likelihood of a passage being identified as fake. Therefore, the assumed high average entropy can serve as an indicator of machine-generated text.

Uniform Information Density (UID) is a statistical metric based on the assumption that humans tend to distribute information uniformly along their text. By analyzing UID-based features, the GPT-who detector [51] captures the unique statistical signature of each author, both human and artificial.

4. Advantages of Feature-Based Methods

Feature-based approaches simplify the understanding of the model’s decision-making process. The discussed techniques, in fact, make the process more transparent and comprehensible by concentrating on particular, quantifiable aspects of the text [72]. Moreover, they are very “flexible” because it is possible to select specific features and adapt the model to particular types of text and writing styles. They are also less computationally intensive, requiring fewer resources and less time than more complex models like deep learning.

5. Limitations of Feature-Based Methods

Despite the numerous existing feature-based models mentioned, there are various issues associated with them that sometimes lead to poor performance. Perturbations (e.g., word substitutions, alterations of characters and words, introduction of spelling errors) can significantly reduce the accuracy of these detectors [5]. Moreover, feature-based models present weaknesses related to the difficulty of transferring specific features between different architectures and sampling methods [48].

C. Neural-Based Methods

In this section, approaches to neural networks are explored by distinguishing between more classical networks and the adoption of pre-trained models or few-shot prompting.

1. Feature-Based Classifiers

Feature-based classifiers can be further differentiated based on the characteristics (features) extracted from the data.

Linguistic Feature-Based Classifiers

When comparing texts generated by large language models (LLMs) with those written by humans, noticing the linguistic differences is crucial to train classifiers that can effectively distinguish them. Text elements can be categorized based on the style of the text, the complexity, the semantic, the psychological, and the knowledge-based characteristics.

These characteristics are extracted mainly by statistical techniques. Subsequently, a classification model can be trained through machine learning techniques [30].

Among the various methods to detect text generated by artificial intelligence, Shah et al. [73] have constructed a classifier based on stylistic features such as frequency analysis of word pairs, language characteristics and lexicographic characteristics. These classifiers based on linguistic characteristics seem to be very beneficial and useful in distinguishing between human-generated and AI-generated texts, but they have flaws that cannot be overlooked: their ability to detect LLM-generated misinformation is limited [5].

Model Feature-Based Classifiers

In addition to the linguistic characteristics, classifiers based on the characteristics of the model have also received considerable attention from research in the field. It is about classifiers that are able to detect

[1] https://gptzero.me/

texts generated by LLMs and trace the origin of the text. In particular, the research made by Su et al. [53] considers the log-rank. However, these methods have a common drawback: they all require access to the source model's logins, so these templates are ineffective when applied to closed sources where the logins are inaccessible.

2. Pre-Training Classifier

Famous pre-learned models, such as Roberta [74], have shown superior performance than traditional machine learning methods and deep learning in text categorization tasks. The 2019 studies identified the improved large language models (LLMs) as Roberta among the best to detect texts generated by other LLMs. These models achieved an average accuracy rate of 95% in their respective fields, surpassing zero-shot and watermarking methods and showing good resistance to different attack techniques. However, like other similar models, these improved encoder-based models are not very robust [54], [75] because they tend to depend too much on training data, leading to a drop in performance with data from different or new domains. Despite this, Roberta-based detectors show remarkable robustness potential, requiring only a few hundred labels to achieve impressive results [55].

3. Fine-Tuning

Fine-tuning, in the field of machine learning and artificial intelligence, is the process of adapting a pre-trained model to a new specific task. Studies of machine-generated text detection have examined how a detection algorithm, such as Roberta, can be trained on a dataset other than that used by an attack model such as GPT-2. It turned out that by perfecting the detection model with only a few hundred samples identified by experts, the detector can greatly improve in adapting to different types of data [55]. This is useful in real situations when a general detector has to deal with a specific attack pattern. When a defender identifies text samples generated by an improved attack model, these examples can be used to make the detection model even more effective [27].

4. Zero-Shot

With the aim of detecting machine-generated text, zero-shot approaches have increasingly become used by researchers and developers. This is related to the fact that zero-shot methods do not need fine-tuning. Some studies show that smaller models of generated text can be used to detect text generated by larger models [57], [58]. This ability decreases as the scale difference grows, while on the contrary, the ability to predict smaller architectures can be very beneficial, as recreating large models with a large number of parameters is highly expensive [27].

5. Advantages of Neural-Based Methods

Neural-based methods are particularly effective in capturing the complex linguistic nuances present in texts by considering specific attributes of advanced models such as ChatGPT [76]. Moreover, fine-tuning can improve the ability to recognize modified texts, making them robust against small mutations in text where even pre-trained models may fail [77].

6. Limitations of Neural-Based Methods

Neural-based methods, generally speaking, need labeled datasets, and their performance and applicability are strictly related to reference domains. Moreover, research findings indicate that the zero-shot approach generally underestimates a simple TF-IDF baseline when attempting to detect output from a generative model that has been developed on a different domain. Because attackers can adjust generative patterns for different purposes, this represents a notable weakness in the zero-shot approach using generative models for detection without tuning [56].

D. Human-Aided Methods

Methods combining features-based or neural techniques with human analysis have been proposed to enhance review capabilities. This integration provides crucial human oversight for trustworthy AI systems but presents scalability challenges due to the need for trained analysts capable of confidently identifying machine-generated text. For example, GLTR (Giant Language Model Test Room) [59] uses a method called "top-k sampling" to highlight words, but this method has been mostly replaced by "nucleus sampling," used in newer models like GPT-3. So, it would probably be difficult for untrained people to detect texts created by the more recent and advanced models. To overcome this limitation, RADAR tester [60] displays the probability each model assigns to a text being AI-generated. A value close to 1 suggests a "high likelihood of AI generation," while a value close to 0 indicates a "high likelihood of human authorship." It also implies that the material is probably produced by a human if the models have significantly different probabilities. Using this information, a human reviewer can effectively assess whether a text was created by a human or an AI.

1. Advantages of Human-Aided Methods

The advantage of these approaches is that they need human support and oversight, which can mitigate the risks associated with a completely autonomous decision-making. This presence also ensure a greater trustworthy in the AI technologies [27].

2. Limitations of Human-Aided Methods

The greatest weakness of Human-Aided Methods is that they are strictly related to the competences (or inabilities) of a human reviewer. Human performance in distinguishing machine-generated text has been extensively studied. Research indicates that untrained individuals often perform no better than chance when distinguishing texts generated by models like GPT-3. However, with some training [61], the accuracy can improve to around 55%.

E. Hybrid Methods

In this section, various hybrid approaches combining and integrating multiple different methodologies to enhance accuracy and reliability in the detection task are explored.

TDA-based detector [65] employs an innovative approach that combines Transformer-based and statistical methodologies to distinguish between human-written and generated texts. This system uses BERT to understand the meaning of words in the context of a text and create detailed representations of them. These representations are then analyzed using Topological Data Analysis (TDA), a mathematical technique that studies the shape and structure of connections between words.

CoCo (Coherence-based Contrastive Learning Model) methodology [66] combines graph-based coherence representation with contrastive learning techniques, aiming to achieve high accuracy in distinguishing between diverse types of textual content. Specifically, CoCo looks at how well the sentences in a text stick together and make sense as a whole (coherence information) and then turns this information into a graph that helps it understand the relationships between different parts of the text. Moreover, contrastive learning helps the model learn better by comparing different texts and focusing on their differences, even under low-resource scenarios.

DIDAN [67] is a tool created to detect fake news articles by analyzing both the text and the images together. It uses a BERT encoder to understand the text and examines the visual-semantic representations to investigate the relationship between the text and images and understand if they match up logically. Additionally, each article is given a score to show how likely it is to be written by a human.

TABLE II. Summary of Used Datasets

Corpus	Adopting Papers	Language	Task
C4 [79]	[34] [36] [68] [35] [36] [69]	English	Language Modelling
RealNews [57]	[48] [65] [66] [67] [37]	English	Text Generation, Language Modelling, Fake News Detection
Webtext [80]	[48] [60] [64] [65]	English	Text Classification, Text Generation, Language Modelling
WikiText-2 [81]	[41] [44] [42]	English, Spanish, German, Swedish	Text Generation, Language Modelling
IMDB [82]	[41] [42] [31]	English	Text Classification, Language Modelling, Paraphrase Identification
AgNews [83]	[41] [31]	English	Text Classification, Zero-Shot Text Classification, Anomaly Detection
SQuAD [84]	[50] [60]	Multilingual	Question Answering, Question Generation
WritingPrompts [85]	[60] [61]	English	Text Generation, Language Modelling, Story Generation, Natural Language Understanding
CoAuthor [86]	[87] [88]	English	Text Generation
PubMed [89]	[49]	Multilingual	Text Summarization, Language Modelling
WMT16 [90]	[50]	English, French, German, Russian,Czech, Finnish, Romanian	Machine Translation
PubMedQA [91]	[50]	English	Question Answering, Language Modelling
RecipeNLG [92]	[61]	Multilingual	Text Generation
Common Crawl [93]	[66]	English	Language Modelling, Generated-Text Detection
NeuralNews [94]	[67]	English	Generated-Text Detection
DialogSum [95]	[32]	English	Text Summarization, Dialogue Generation, Abstractive Text Summarization
DBpedia [96]	[36]	English	Text Classification
HC3 [97]	[43]	English, Chinese	Text Classification, Question Answering, Sentence Similarity, Zero-Shot Classification
GLUE [98]	[31]	English	Text Classification, Natural Language Inference, Semantic Textual Similarity, Natural Language Understanding, Semantic Textual Similarity within Bi-Encoder
SNLI [99]	[31]	English	Natural Language Inference

1. Advantages of Hybrid Methods

Modern and advanced hybrid approaches for authorship attribution, which combine multiple methods, make it more complicated and challenging for both human authors and LLMs to obfuscate their style or deceive detection systems, especially when it comes to artificially generated texts [30]. Moreover, by leveraging both traditional and new technologies, detectors can benefit from different strengths, related to each specific component, and give exhaustive results [78].

2. Limitations of Hybrid Methods

The main problem related to Hybrid approaches is that, requiring the integration of multiple models, the overall complexity inevitably increases, necessitating of considerable computing power and memory resources. This reflects a greater issue of scalability and the need for optimization for large volumes of data.

VI. Detection Datasets

From the literature analysis, emerging frequently adopted datasets are ones grouped in Table II. They are not all strictly related to the machine-generated text detection task but are derived from different natural language processing tasks. In many analyzed works, in fact, existing datasets are adopted as examples of human-written text, while machine-generated text is produced ad-hoc by a selected LLM. This means that learning models' capacity to identify generated content is often related to the application domain of the adopted dataset and the LLM adopted to produce new text. Another important limitation concerns the fact that the majority of the corpus is written in English. This means that constructed models are more powerful in detecting generated text in English.

VII. Detection Tools

Considered literature also includes the performance evaluation of tools for machine-generated text detection available at the state-of-the-art level. This section tries to summarize their results in order to highlight possible practical solutions to apply or provide a benchmark for new implementations. The section also investigates how the detectors have been developed.

Copyleaks[2] combines many techniques. Trillions of data were collected from universities and enterprises worldwide to train the model. TurnItIn[3] model is trained on AI-generated and academic writing. They also gave significance to the language they considered. Indeed, they included second language learners or texts written by people who use English but who are not native speakers. The training data is based on different subject areas. Scribbr[4] uses the analysis of stylistic patterns and sentence structure, and it employs algorithms that have been trained on big collections of content written by humans

[2] https://copyleaks.com/ai-content-detector

[3] https://www.turnitin.com/

[4] https://www.scribbr.com/ai-detector/

TABLE III. Performance of Detection Tools

Detector	Accuracy [100]	Accuracy [101]	Fee	Overview
Copyleaks	100%	91%	Free with limitations	Details about the model are not publicly available.
Turnitin	100%	-	Institutional subscription	The model is trained to detect wordprobability differences, leveraging the principle that AI generates words predictably, while human writing is more varied and unpredictable.
Originality.ai	98%	-	$0.01 per 100 words	The tool uses supervised learning with several models, including BERT and a version of Roberta.
Scribbr	88%	-	Free with limitations	The tool uses the analysis of stylistic pattern and sentence structure. It employs algorithms that have been trained on big collections of content written by humans and generated by machines.
ZeroGPT	87%	-	Free with limitations	The tool utilizes a multi-stage deep learning methodology (*DeepAnalyse*) developed by the ZeroGPT's team and trained on different kinds of datasets.
Writer	71%	99%	Free with limitations	Details about the model are not publicly available.
Content at Scale	71%	48%	Free with limitations	The tool uses both NLP and a trained model to identify specific aspects that lead to a higher likelihood of a text being detected as AI-generated (e.g., predicting likely next word choices and recognizing sentence structure).

and generated by machines. Originality.ai[5] uses supervised learning with several models, including BERT and a version of Roberta, and it has been trained on millions of examples of generated and human text. ZeroGPT[6] utilizes the so-called DeepAnalyse technology (a multi-stage methodology developed by ZeroGPT's team) to determine the origin of a given text, leveraging a deep learning methodology trained on different kinds of datasets. Content at Scale[7] uses both natural language processing and a trained model to identify specific aspects that lead to a higher likelihood of a text being detected as AI-generated, for example, by predicting likely next-word choices and recognizing sentence structure. Details about the model used by Writer[8] are not publicly available.

Table III contains a summary of aforementioned tools by reporting their performance in terms of Accuracy [100] [101]. It highlights interesting performance even for free solutions. Nevertheless, one must take into account that, in general, the performance of available tools decreases with the adoption of GPT-4 [102] or by paraphrasing the text [103].

VIII. Discussion

Reviewing the existing methodologies for distinguishing between human-written and machine-generated texts revealed some ongoing problems and limitations. These aspects must be taken into account when developing new discriminators or methodologies to enhance the effectiveness of proposed solutions.

A. Languages

Current methodologies may perform well for English texts but may not produce the same results for other languages [104].

The variation in grammar, syntax, and idiomatic expressions presents a significant challenge. Findings indicate that most existing black-box methods are ineffective when applied in multilingual environments, with statistical approaches significantly trailing behind fine-tuned models [105]. So, it is crucial to develop approaches that are effective across a wide range of languages [106].

B. Hybrid Text

With the increase in hybrid texts, which combine human content with generated content, the analysis of such texts becomes more complex. Current methodologies may not be robust enough to handle this complexity, necessitating adaptations or new strategies specifically for hybrid texts [87]. Zeng et al. [88] highlight that generated-text detection with hybrid texts is tough for several reasons:

1. Human writers often select and edit machine-generated phrases based on their personal style;
2. The swap of authorship between adjacent sentences creates difficulties for segment detectors;
3. Brief text segments give little stylistic evidence, not allowing definitive authorship identification.

C. New-Generation LLMs

New large language models are rapidly developing, bringing new capabilities and challenges. Continuously testing and updating discrimination methods is essential to ensure that evaluations remain accurate and relevant because methodologies that worked well with previous models may no longer be effective [107].

D. Lack of Regulation

The application of regulations by LLM developers would be useful in making the generated text a more reliable tool and less prone to misuse. In particular, measures like the *AI Act* [108] and the *Telephone Consumer Protection Act* [109] aiming to regulate artificial intelligence adoption should be fine-tuned.

IX. Conclusions

Despite the significant progress in the generated-text detection task, there are still many challenges to address. Indeed, the effectiveness of current detection methodologies is strictly related to the application context and the complexity of the text under analysis. Recent methods, such as those using Transformer models, show promising results in terms of accuracy but require many computational resources and may be less effective for multilingual contexts or short texts. In contrast, traditional statistical methods need fewer resources but suffer from limitations in terms of precision and adaptability to complex texts.

[5] https://originality.ai/ai-checker

[6] https://www.zerogpt.com/

[7] https://contentatscale.ai/ai-content-detector/

[8] https://writer.com/ai-content-detector/

For these reasons, hybrid approaches, which combine different methodologies, could be the most promising solution. However, the rise of hybrid texts, blending human-generated content with machine-generated content, presents new challenges that require further research and innovation to overcome.

The lack of specific regulations regarding generated text is a major hurdle to overcome. The existence of rules would help ensure the recognition of generated text and make users more conscious of what they are reading. Therefore, regulations should be involved in the generated text detection task.

Acknowledgment

This work was partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

References

[1] F. García-Peñalvo, A. Vázquez-Ingelmo, *et al.*, "What do we mean by genai? a systematic mapping of the evolution, trends, and techniques involved in generative ai," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4, pp. 7–16, 2023.

[2] J. Oluwaseyi, K. Potter, "Exploring natural language generation (nlg) methods for generating human-like text from structured or unstructured data," *Journal of Machine to Machine Communications*, 12 2023.

[3] A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, B. Agyemang, "What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education," *Smart learning environments*, vol. 10, no. 1, p. 15, 2023.

[4] M. Alier, F. J. García-Peñalvo, J. D. Camba, "Generative Artificial Intelligence in education: From deceptive to disruptive," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 5, Special issue on Generative Artificial Intelligence in Education, pp. 5–14, 2024.

[5] J. Wu, S. Yang, R. Zhan, Y. Yuan, D. F. Wong, L. S. Chao, "A survey on llm-generated text detection: Necessity, methods, and future directions," *ArXiv*, vol. abs/2310.14724, 2023.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, Red Hook, NY, USA, 2017, p. 6000–6010, Curran Associates Inc.

[7] Z. Yang, "Xlnet: Generalized autoregressive pretraining for language understanding," *arXiv preprint arXiv:1906.08237*, 2019.

[8] S. Grassini, "Shaping the future of education: exploring the potential and consequences of ai and chatgpt in educational settings," *Education Sciences*, vol. 13, no. 7, p. 692, 2023.

[9] J. Jeon, S. Lee, "Large language models in education: A focus on the complementary relationship between human teachers and chatgpt," *Education and Information Technologies*, 05 2023, doi: 10.1007/s10639-023-11834-1.

[10] V. Parra, P. Sureda, A. Corica, S. Schiaffino, D. Godoy, "Can Generative AI solve Geometry Problems? Strengths and Weaknesses of LLMs for Geometric Reasoning in Spanish," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 5, Special issue on Generative Artificial Intelligence in Education, pp. 65–74, 2024, doi: 10.9781/ijimai.2024.02.009.

[11] A. Thirunavukarasu, D. Ting, K. Elangovan, L. Gutierrez Sinisterra, T. Tan, D. Ting, "Large language models in medicine," *Nature Medicine*, vol. 29, 07 2023, doi: 10.1038/s41591-023-02448-8.

[12] S. Hajijama, D. Juneja, P. Nasa, "Large language model in critical care medicine: Opportunities and challenges," *Indian Journal of Critical Care Medicine*, vol. 28, no. 6, pp. 523–525, 2024.

[13] Y. Feng, S. Vanam, M. Cherukupally, W. Zheng, M. Qiu, H. Chen, "Investigating code generation performance of chatgpt with crowdsourcing social data," in *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2023, pp. 876–885.

[14] R. Khoury, A. R. Avila, J. Brunelle, B. M. Camara, "How secure is code generated by chatgpt?," in *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2023, pp. 2445–2451.

[15] V. C. C. Kwok-Yan Lam, Z. K. Yeong, "Applying large language models for enhancing contract drafting," in *Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workspace (LegalAIIA 2023)*, 2023.

[16] J. H. Choi, "How to use large language models for empirical legal research," *Journal of Institutional and Theoretical Economics (Forthcoming)*, 2023.

[17] J. Cui, Z. Li, Y. Yan, B. Chen, L. Yuan, "Chatlaw: Open- source legal large language model with integrated external knowledge bases," *arXiv preprint arXiv:2306.16092*, 2023.

[18] J. Tan, H. Westermann, K. Benyekhlef, "Chatgpt as an artificial lawyer?," in *AI4AJ@ ICAIL*, 2023.

[19] X. Fang, S. Che, M. Mao, H. Zhang, M. Zhao, X. Zhao, "Bias of ai-generated content: an examination of news produced by large language models," *Scientific Reports*, vol. 14, no. 1, p. 5224, 2024.

[20] C. Novelli, F. Casolari, P. Hacker, G. Spedicato, L. Floridi, "Generative ai in eu law: Liability, privacy, intellectual property, and cybersecurity," *EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity (January 14, 2024)*, 2024.

[21] L. Tang, Y.-S. Su, "Ethical Implications and Principles of Using Artificial Intelligence Models in the Classroom: A Systematic Literature Review," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 8, no. 5, 2024.

[22] I. Ulnicane, "Governance fix? power and politics in controversies about governing generative ai," *Policy and Society*, p. puae022, 2024.

[23] S. S. Ghosal, S. Chakraborty, J. Geiping, F. Huang, D. Manocha, A. Bedi, "A survey on the possibilities & impossibilities of ai-generated text detection," *Transactions on Machine Learning Research*, 2023.

[24] D. Beresneva, "Computer-generated text detection using machine learning: A systematic review," in *Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings 21*, 2016, pp. 421–426, Springer.

[25] G. Jawahar, M. Abdul-Mageed, V. Laks Lakshmanan, "Automatic detection of machine generated text: A critical survey," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 2296–2309.

[26] M. Dhaini, W. Poelman, E. Erdogan, "Detecting chatgpt: A survey of the state of detecting chatgpt-generated text," in *Proceedings of the 8th Student Research Workshop associated with the International Conference Recent Advances in Natural Language Processing*, 2023, pp. 1–12.

[27] E. Crothers, N. Japkowicz, H. Viktor, "Machine-generated text: A comprehensive survey of threat models and detection methods," *IEEE Access*, vol. PP, pp. 1–1, 01 2023, doi: 10.1109/ACCESS.2023.3294090.

[28] R. Tang, Y.-N. Chuang, X. Hu, "The science of detecting llm-generated text," *Communications of the ACM*, vol. 67, p. 50–59, mar 2024, doi: 10.1145/3624725.

[29] X. Yang, L. Pan, X. Zhao, H. Chen, L. R. Petzold, W. Y. Wang, W. Cheng, "A survey on detection of llms-generated content," *ArXiv*, vol. abs/2310.15654, 2023.

[30] A. Uchendu, T. Le, D. Lee, "Attribution and obfuscation of neural text authorship: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 25, no. 1, pp. 1–18, 2023.

[31] Chenxi Gu and Chengsong Huang and Xiaoqing Zheng and Kai-Wei Chang and Cho-Jui Hsieh, "Watermarking Pre-trained Language Models with Backdooring," *ArXiv*, vol. abs/2210.07543, 2022.

[32] E. Lucas, T. Havens, "GPTs don't keep secrets: Searching for backdoor watermark triggers in autoregressive language models," in *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, Toronto, Canada, July 2023, pp. 242–248, Association for Computational Linguistics.

[33] R. Tang, Q. Feng, N. Liu, F. Yang, X. Hu, "Did you train on my dataset? towards public dataset protection with cleanlabel backdoor watermarking," *ACM SIGKDD Explorations Newsletter*, vol. 25, p. 43–53, jul 2023, doi: 10.1145/3606274.3606279.

[34] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, T. Goldstein, "A watermark for large language models," in *Proceedings of the 40th International Conference on Machine Learning*, vol. 202 of *Proceedings of Machine Learning Research*, 23–29 Jul 2023, pp. 17061–17084, PMLR.

[35] J. Kirchenbauer, J. Geiping, Y. Wen, M. Shu, K. Saifullah, K. Kong, K. Fernando, A. Saha, M. Goldblum, T. Goldstein, "On the reliability of watermarks for large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[36] A. Liu, L. Pan, X. Hu, S. Li, L. Wen, I. King, S. Y. Philip, "An unforgeable publicly verifiable watermark for large language models," in *The Twelfth International Conference on Learning Representations*, 2023.

[37] A. Hou, J. Zhang, T. He, Y. Wang, Y.-S. Chuang, H. Wang, L. Shen, B. Van Durme, D. Khashabi, Y. Tsvetkov, "Semstamp: A semantic watermark with paraphrastic robustness for text generation," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 4067–4082.

[38] Por, Lip Yee and Wong, KokSheik and Chee, Kok Onn, "UniSpaCh: A text-based data hiding method using Unicode space characters," *Journal of Systems and Software*, vol. 85, pp. 1075–1082, may 2012, doi: 10.1016/j.jss.2011.12.023.

[39] U. Topkara, M. Topkara, M. J. Atallah, "The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions," in *Proceedings of the 8th workshop on Multimedia and security*, 2006, pp. 164–174.

[40] Munyer, Travis and Tanvir, Abdullah and Das, Arjon and Zhong, Xin, "DeepTextMark: A Deep Learning- Driven Text Watermarking Approach for Identifying Large Language Model Generated Text," *IEEE Access*, vol. PP, pp. 1–1, 01 2024, doi: 10.1109/ACCESS.2024.3376693.

[41] X. Yang, J. Zhang, K. Chen, W. Zhang, Z. Ma, F. Wang, N. Yu, "Tracing text provenance via context-aware lexical substitution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 11613–11621.

[42] K. Yoo, W. Ahn, J. Jang, N. Kwak, "Robust multi-bit natural language watermarking through invariant features," in *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

[43] X. Yang, K. Chen, W. Zhang, C. Liu, Y. Qi, J. Zhang, H. Fang, N. Yu, "Watermarking text generated by black-box language models," *arXiv preprint arXiv:2305.08883*, 2023.

[44] S. Abdelnabi, M. Fritz, "Adversarial watermarking transformer: Towards tracing text provenance with data hiding," in *2021 IEEE Symposium on Security and Privacy (SP)*, 2021, pp. 121–140, IEEE.

[45] A. Modupe, T. Celik, V. Marivate, O. Olugbara, "Post- authorship attribution using regularized deep neural network," *Applied Sciences*, vol. 12, p. 7518, 07 2022, doi: 10.3390/app12157518.

[46] T. Schuster, R. Schuster, D. Shah, R. Barzilay, "The limitations of stylometry for detecting machine-generated fake news," *Computational Linguistics*, vol. 46, pp. 1–18, 03 2020, doi: 10.1162/COLI_a_00380.

[47] T. Kumarage, J. Garland, A. Bhattacharjee, K. Trapeznikov, S. Ruston, H. Liu, "Stylometric detection of ai-generated text in twitter timelines," *arXiv preprint arXiv:2303.03697*, 2023.

[48] L. Frohling, A. Zubiaga, "Feature-based detection of automated language models: tackling gpt-2, gpt-3 and grover," *PeerJ Computer Science*, vol. 7, p. e443, 04 2021, doi: 10.7717/peerj-cs.443.

[49] A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, T. Goldstein, "Spotting llms with binoculars: Zero-shot detection of machine- generated text," *arXiv preprint arXiv:2401.12070*, 2024.

[50] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, "Detectgpt: zero-shot machine-generated text detection using probability curvature," in *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, 2023, JMLR.org.

[51] S. Venkatraman, A. Uchendu, D. Lee, "Gpt-who: An information density-based machine-generated text detector," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 103–115.

[52] A. Aich, S. Bhattacharya, N. Parde, "Demystifying neural fake news via linguistic feature-based interpretation," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 6586–6599.

[53] J. Su, T. Zhuo, D. Wang, P. Nakov, "Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 12395–12412.

[54] A. Bakhtin, S. Gross, M. Ott, Y. Deng, M. Ranzato, A. Szlam, "Real or fake? learning to discriminate machine from human generated text," *arXiv preprint arXiv:1906.03351*, 2019.

[55] J. D. Rodriguez, T. Hay, D. Gros, Z. Shamsi, R. Srinivasan, "Cross-domain detection of gpt-2-generated technical text," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: human language technologies*, 2022, pp. 1213–1233.

[56] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert- Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, *et al.*, "Release strategies and the social impacts of language models," *arXiv preprint arXiv:1908.09203*, 2019.

[57] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, Y. Choi, "Defending against neural fake news," *Advances in neural information processing systems*, vol. 32, 2019.

[58] E. Crothers, N. Japkowicz, H. Viktor, P. Branco, "Adversarial robustness of neural-statistical features in detection of generative transformers," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–8, IEEE.

[59] S. Gehrmann, H. Strobelt, A. M. Rush, "Gltr: Statistical detection and visualization of generated text," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2019, pp. 111–116.

[60] X. Hu, P.-Y. Chen, T.-Y. Ho, "Radar: Robust ai-text detection via adversarial learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 15077–15095, 2023.

[61] E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, N. A. Smith, "All that's 'human'is not gold: Evaluating human evaluation of generated text," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 7282–7296.

[62] D. Ippolito, D. Duckworth, D. Eck, "Automatic detection of generated text is easiest when humans are fooled," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1808–1822.

[63] A. Uchendu, Z. Ma, T. Le, R. Zhang, D. Lee, "Turingbench: A benchmark environment for turing test in the age of neural text generation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 2001–2016.

[64] Y. Dou, M. Forbes, R. Koncel-Kedziorski, N. A. Smith, Y. Choi, "Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 7250–7274.

[65] L. Kushnareva, D. Cherniavskii, V. Mikhailov, E. Artemova, S. Barannikov, A. Bernstein, I. Piontkovskaya, D. Piontkovski, E. Burnaev, "Artificial text detection via examining the topology of attention maps," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 635–649.

[66] X. Liu, Z. Zhang, Y. Wang, H. Pu, Y. Lan, C. Shen, "Coco: Coherence-enhanced machine-generated text detection under low resource with contrastive learning," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 16167–16188.

[67] R. Tan, B. Plummer, K. Saenko, "Detecting cross- modal inconsistency to defend against neural fake news," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 2081–2106.

[68] L. Wang, W. Yang, D. Chen, H. Zhou, Y. Lin, F. Meng, J. Zhou, X. Sun, "Towards codable text watermarking for large language models," *arXiv preprint arXiv:2307.15992*, 2023.

[69] R. Kuditipudi, J. Thickstun, T. Hashimoto, P. Liang "Robust Distortion-free Watermarks for Language Models," *Transactions on Machine Learning Research*, 2024.

[70] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre- training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019, Association for Computational Linguistics.

[71] A. Muñoz-Ortiz, C. Gómez-Rodríguez, D. Vilares, "Contrasting linguistic patterns in human and llm- generated text," *arXiv preprint arXiv:2308.09067*, 2023.

[72] R. Corizzo, S. Leal-Arenas, "One-class learning for ai- generated essay detection," *Applied Sciences*, vol. 13, no. 13, 2023, doi: 10.3390/app13137901.

[73] A. Shah, P. Ranka, U. Dedhia, S. Prasad, S. Muni, K. Bhowmick, “Detecting and unmasking ai-generated texts through explainable artificial intelligence using stylistic features,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 10, 2023.

[74] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.

[75] D. Macko, R. Moro, A. Uchendu, J. Lucas, M. Yamashita, M. Pikuliak, I. Srba, T. Le, D. Lee, J. Simko, *et al.*, “Multitude: Large-scale multilingual machine-generated text detection benchmark,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 9960–9987.

[76] R. Gaggar, A. Bhagchandani, H. Oza, “Machine-generated text detection using deep learning,” 2023. [Online]. Available: https://arxiv.org/abs/2311.15425.

[77] J. A. Guerrero, “Detecting ai generated text using neural networks,” Master’s thesis, Texas A&M University, 2023.

[78] Y. Zhang, Q. Leng, M. Zhu, R. Ding, Y. Wu, J. Song, Y. Gong, “Enhancing text authenticity: A novel hybrid approach for ai-generated text detection,” in *2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI)*, 2024, pp. 433–438.

[79] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[80] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[81] S. Merity, C. Xiong, J. Bradbury, R. Socher, “Pointer sentinel mixture models,” *arXiv preprint arXiv:1609.07843*, 2016.

[82] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, C. Potts, “Learning word vectors for sentiment analysis,” in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011, pp. 142–150.

[83] X. Zhang, J. Zhao, Y. LeCun, “Character-level convolutional networks for text classification,” *Advances in neural information processing systems*, vol. 28, 2015.

[84] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, “Squad: 100,000+ questions for machine comprehension of text,” *arXiv preprint arXiv:1606.05250*, 2016.

[85] A. Fan, M. Lewis, Y. Dauphin, “Hierarchical neural story generation,” *arXiv preprint arXiv:1805.04833*, 2018.

[86] M. Lee, P. Liang, Q. Yang, “Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities,” in *Proceedings of the 2022 CHI conference on human factors in computing systems*, 2022, pp. 1–19.

[87] A. Richburg, C. Bao, M. Carpuat, “Automatic authorship analysis in human-ai collaborative writing,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC- COLING 2024)*, 2024, pp. 1845–1855.

[88] Z. Zeng, S. Liu, L. Sha, Z. Li, K. Yang, S. Liu, D. Gavsevi’c, G. Chen, “Detecting ai-generated sentences in human- ai collaborative hybrid texts: Challenges, strategies, and insights,” *arXiv preprint arXiv:2403.03506v4*, 2024.

[89] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, T. Eliassi-Rad, “Collective classification in network data,” *AI magazine*, vol. 29, no. 3, pp. 93–93, 2008.

[90] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, V. Logacheva, C. Monz, *et al.*, “Findings of the 2016 conference on machine translation (wmt16),” in *First conference on machine translation*, 2016, pp. 131–198, Association for Computational Linguistics.

[91] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, X. Lu, “Pubmedqa: A dataset for biomedical research question answering,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP- IJCNLP)*, 2019, pp. 2567–2577.

[92] M. Bień, M. Gilski, M. Maciejewska, W. Taisner, D. Wisniewski, A. Lawrynowicz, “Recipenlg: A cooking recipes dataset for semi-structured text generation,” in *Proceedings of the 13th International Conference on Natural Language Generation*, 2020, pp. 22–28.

[93] J. M. Patel, J. M. Patel, “Introduction to common crawl datasets,” *Getting structured data from the internet: running web crawlers/scrapers on a big data production scale*, pp. 277–324, 2020.

[94] R. Tan, B. Plummer, K. Saenko, “Detecting cross- modal inconsistency to defend against neural fake news,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 2081–2106.

[95] Y. Chen, Y. Liu, L. Chen, Y. Zhang, “Dialogsum: A real-life scenario dialogue summarization dataset,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 5062–5074.

[96] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, “Dbpedia: A nucleus for a web of open data,” in *international semantic web conference*, 2007, pp. 722–735, Springer.

[97] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, Y. Wu, “How close is chatgpt to human experts? comparison corpus, evaluation, and detection,” *arXiv preprint arXiv:2301.07597*, 2023.

[98] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, “Glue: A multi-task benchmark and analysis platform for natural language understanding,” in *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[99] S. Bowman, G. Angeli, C. Potts, C. D. Manning, “A large annotated corpus for learning natural language inference,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 632–642.

[100] W. H. Walters, “The effectiveness of software designed to detect ai-generated writing: A comparison of 16 ai text detectors,” *Open Information Science*, vol. 7, no. 1, p. 20220158, 2023.

[101] N. Ladha, K. Yadav, P. Rathore, “Ai-generated content detectors: Boon or bane for scientific writing,” *Indian Journal of Science and Technology*, vol. 16, no. 39, pp. 3435–3439, 2023.

[102] G.-A. Odri, D. J. Y. Yoon, “Detecting generative artificial intelligence in scientific articles: evasion techniques and implications for scientific integrity,” *Orthopaedics & Traumatology: Surgery & Research*, vol. 109, no. 8, p. 103706, 2023.

[103] M. Perkins, J. Roe, B. H. Vu, D. Postma, D. Hickerson, J. McGaughran, H. Q. Khuat, “Genai detection tools, adversarial techniques and implications for inclusivity in higher education,” *arXiv preprint arXiv:2403.19148*, 2024.

[104] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, J. Zou, “Gpt detectors are biased against non-native english writers,” *Patterns*, vol. 4, no. 7, p. 100779, 2023, doi: https://doi.org/10.1016/j.patter.2023.100779.

[105] D. Macko, R. Moro, A. Uchendu, J. Lucas, M. Yamashita, M. Pikuliak, I. Srba, T. Le, D. Lee, J. Simko, M. Bielikova, “MULTITuDE: Large-scale multilingual machine- generated text detection benchmark,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, Dec. 2023, pp. 9960–9987, Association for Computational Linguistics.

[106] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, O. M. Afzal, T. Mahmoud, G. Puccetti, T. Arnold, *et al.*, “Semeval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection,” *arXiv preprint arXiv:2404.14183*, 2024.

[107] A. Bhattacharjee, R. Moraffah, J. Garland, H. Liu, “Eagle: A domain generalization framework for ai-generated text detection,” *arXiv preprint arXiv:2403.15690*, 2024.

[108] O. J. of the European Union, “Regulation (eu) 2024/1689 of the european parliament and of the council.” https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689. Last accessed: 23 September 2024.

[109] F. C. Commission, “Telephone consumer protection act.” https://docs.fcc.gov/public/attachments/ DOC-404036A1.pdf. Last accessed: 23 September 2024.

Serena Fariello

Serena Fariello received a bachelor’s degree in Statistics for Big Data from the University of Salerno, Italy, in 2023. She is currently a Data Science and Innovation Management student at the same university.

Giuseppe Fenza

Giuseppe Fenza received the Graduate degree and the Ph.D. degree in computer sciences from the University of Salerno, Italy, in 2004 and 2009, respectively. He is currently an Associate Professor in computer science with the University of Salerno. He has over 60 publications in fuzzy decision making, knowledge extraction and management, situation and context awareness, semantic information retrieval, service oriented architecture, and ontology learning. More recently, he has worked in automating open source intelligence and big data analytics for counterfeiting extremism and supporting information disorder awareness. His research interests include computational intelligence methods to support semantic-enabled solutions and decision-making.

Flavia Forte

Flavia Forte received a bachelor's degree in Computer Engineering from the University of Salerno, Italy, in 2023. She is currently a Data Science and Innovation Management student at the same university.

Mariacristina Gallo

Mariacristina Gallo received her master's degree in computer science and Ph.D. in Big Data Management from the University of Salerno, Italy, in 2009 and 2021, respectively. Since 2022, she has been an Assistant Professor at the University of Salerno, Italy, where she works mainly on Social Media Analytics, Open Source Intelligence, Machine Learning, Deep Learning, Explainable Artificial Intelligence, Large Language Models applied to domains such as IoT, Environmental Security and Information Disorder.

Martina Marotta

Martina Marotta received a bachelor's degree in Computer Engineering from the University of Salerno, Italy, in 2023. She is currently a Data Science and Innovation Management student at the same university.

Regular Issue

Large Language Models for in Situ Knowledge Documentation and Access With Augmented Reality

Juan Izquierdo-Domenech, Jordi Linares-Pellicer, Isabel Ferri-Molla *

Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València (UPV), València (Spain)

* Corresponding author: juaizdom@upv.es (J. Izquierdo-Domenech), jlinares@dsic.upv.es (J. Linares-Pellicer), isfermol@upv.es (I. Ferri-Molla).

Received 16 February 2023 | Accepted 22 June 2023 | Published 29 September 2023

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

Augmented reality (AR) has become a powerful tool for assisting operators in complex environments, such as shop floors, laboratories, and industrial settings. By displaying synthetic visual elements anchored in real environments and providing information for specific tasks, AR helps to improve efficiency and accuracy. However, a common bottleneck in these environments is introducing all necessary information, which often requires predefined structured formats and needs more ability for multimodal and Natural Language (NL) interaction. This work proposes a new method for dynamically documenting complex environments using AR in a multimodal, non-structured, and interactive manner. Our method employs Large Language Models (LLMs) to allow experts to describe elements from the real environment in NL and select corresponding AR elements in a dynamic and iterative process. This enables a more natural and flexible way of introducing information, allowing experts to describe the environment in their own words rather than being constrained by a predetermined structure. Any operator can then ask about any aspect of the environment in NL to receive a response and visual guidance from the AR system, thus allowing for a more natural and flexible way of introducing and retrieving information. These capabilities ultimately improve the effectiveness and efficiency of tasks in complex environments.

Keywords

Augmented Reality,
Deep Learning,
Multimodal Interaction,
Large Language Models,
Transformers.

DOI: 10.9781/ijimai.2023.09.002

I. Introduction

Augmented Reality (AR) and its capability for superimposing synthetic elements on top of real environments has been, indeed, a key factor in the rise of Industry 4.0 [1]. There are numerous definitions of AR, with one of the most well-known being the one proposed by Azuma: *"AR is a system that supplements the real world with virtual (computer-generated) objects that appear to coexist in the same space as the real world"* [2]. Billinghurst also defines AR as an interactive experience in which real-world objects are enhanced by computer-generated perceptual information [3]; nonetheless, AR had been applied in the industry field even before such definitions [4]. The ability to enhance environments with this technology has been utilized in various industrial applications, including product design, process design and control, maintenance processes, and learning. Its benefits have been widely demonstrated. Examples include using AR to visualize and manipulate 3D models during the design process, to provide real-time guidance and instructions for Maintenance, Repair, and Overhaul (MRO) tasks, and to enhance training and education programs through interactive simulations and visualizations. Industry 4.0 lays on several pillars, such as the Industrial Internet of Things (IIoT), cloud computing, additive manufacturing, and AR; however, the latter is unique in that it focuses on the human factor [5]. On the other hand, shop floor operators have seen how their roles and required knowledge have been transformed to match completely different profiles, leading to a need for more skilled operators with advanced education in the use of technologies [6]–[9]. AR can serve as an assistive technology to support shop floor operators in these environments.

The evolution of Artificial Intelligence (AI) and its integration into the industry is one the most critical components behind what it has been defined as Industry 4.0. It can lead to a shift in the role of workers towards more value-added tasks, which can increase job satisfaction and improve overall productivity. By incorporating these technologies, manufacturers can create a more efficient and flexible workforce, ultimately leading to a better future of work in the manufacturing industry [10]. The current proposal is a step forward in achieving these objectives.

Please cite this article in press as:
J. Izquierdo-Domenech, J. Linares-Pellicer, I. Ferri-Molla. Large Language Models for in Situ Knowledge Documentation and Access With Augmented Reality, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 19-29, 2025, http://dx.doi.org/10.9781/ijimai.2023.09.002

The proper training of operators is always the first challenge to be met to guarantee their subsequent ability to work effectively and efficiently. Apart from the emergence of new possibilities in this training, such as multimedia tools, Virtual Reality (VR) and AR, 'one-to-one' training is still very beneficial. Direct interaction is still a precious element in training in complex contexts, such as laboratories, control centers, and shop floors in the industry. However, after training, operators need immediate access to documentation that can solve new doubts or problems that may arise at any time. On these occasions, the presence of a specialized expert is very rare or impossible. In contrast to initial training, it is unlikely that the expert or Subject Matter Expert (SME) will be available for the operator's day-to-day work [11].

It is, therefore, essential to develop solutions that enable the operator to access information quickly and efficiently in case of need. In this context, there is a need to know what means and interrogation mechanisms are available. An ideal solution would offer multiple interaction options, including operators' ability to ask questions and receive answers in Natural Language (NL), as discussed in [12]. It is necessary to consider technical documentation and expert knowledge to provide adequate answers. It is very interesting to offer not only the possibility to give answers in NL to the operator based on the technical documentation, but also the information provided by the experts; however, technical documentation and expert information are typically unstructured, presenting a significant challenge for creating operator assistance systems in complex environments. This often leads to the creation of time-consuming, *ad hoc* solutions for different environments, which can be overwhelming and cause an excessive workload, specially in environments with a high degree of diversity or work volume. Therefore, it is necessary to address the issue of unstructured technical documentation and expert information to create operator assistance systems that are efficient, effective, and sustainable in complex environments.

One of the latest technologies based on Deep Learning (DL), Large Language Models (LLMs), can aid in NL interaction and information retrieval by operators. The proposed system enables experts to train operators on the job, allowing for the system to serve as a knowledge source for subsequent consultations. This type of learning, known as in situ learning or Scenario Based Training (SBT), has been demonstrated to enhance knowledge acquisition and retention according to prior research [13].

One-on-one training, primarily SBT, is still the best way to provide knowledge of complex systems. Combining SBT with an automatic acquisition of information, adding multimodal elements, avoiding the need to structure or post-process the information, and making this knowledge available to the operators, is an element of great interest, and the main interest of this work.

Another issue when discussing complex environments, such as shop floors, is documentation access. The complexity of these environments tends to increase exponentially, as does the specialized knowledge and technology required by operators. Documentation about the different machines spread over a shop floor is critical for making them work and learning and maintenance processes. However, traditional forms of documentation, such as paper manuals, can be cumbersome and difficult to use due to their lack of portability, the potential for inaccuracies, and interpretation issues. As Ventura highlights, these issues can make it difficult to effectively utilize this type of information [14]. To address these challenges, several alternatives using AR technology have been proposed. For example, AR can be used to display machine-specific documentation on a user device in real-time as they work, allowing for easier reference and reducing the risk of errors due to outdated or incomplete information [8], [15], [16]. In the context of Industry 4.0, the need for such accessible and reliable information becomes even more pressing as the demands for decentralized, accurate, modular, and fast access to information become increasingly important [17]. By utilizing AR technology, it may be possible to improve the accessibility and usability of documentation in complex environments such as shop floors, ultimately leading to increased efficiency and productivity.

AI and its subsets, such as Machine Learning (ML) and DL, also play an indispensable role in the industry 4.0 field. The capabilities to develop solutions that range from Computer Vision (CV), Natural Language Processing (NLP), and finding patterns hidden in vast amounts of data are being applied in tasks such as predictive maintenance [18], process automation [19], or security enhancement [20]. Some architectures that enable developing applications that tackle these kinds of tasks are Convolutional Neural Networks (CNNs) for CV and Transformers for NLP. CNNs are a type of neural network architecture typically used for image classification, while Transformers have revolutionized NLP tasks by allowing for attention-based mechanisms to capture semantic dependencies between words. Transformers are behind the current LLMs and are mainly used for tasks such as language translation, text summarisation, sentiment analysis, question answering (QA), and language modeling [21]. AI tools are used to analyze large amounts of data, automate repetitive tasks, and improve decision-making processes, leading to increased efficiency, cost savings, and improved customer experiences. AI-powered systems are also being used to monitor, predict, and prevent potential equipment failures and downtime, reducing maintenance costs and increasing overall productivity.

This study aims to evaluate the effectiveness of using multimodal interaction and AR to enrich complex environments with additional information from a variety of sources (e.g., technical documentation, experience-based knowledge) in an unstructured format and to assess the feasibility of novice workshop operators accessing this information multimodally by anchoring it in the environment through AR.

The main contributions of this work are:

1. The ability to incorporate the knowledge and experience of SMEs in a flexible format and to continually update it through an iterative process,
2. The collection of information in multiple formats, anchored in the physical space and using NL, to reduce the need for access to technical documentation and SMEs,
3. Reducing the time between the emergence of a doubt and receiving a response.

This paper is structured as follows: In section II, numerous examples of AR and AI being applied to industrial settings are enumerated and described to emphasize this research's novelties. Then, in section III, the principal user roles and technology used for this research are explained. Section IV presents and evaluates the results from the experiment. Finally, in section V, the conclusions are described, highlighting the significance of the findings.

II. Background and Context

Technological advancements have led to the integration of AR and AI into various industries in recent years. These technologies have the potential to revolutionize the way shop floor operators and SMEs interact with and perceive the environment around them. This section will explore some of the current state-of-the-art applications of AR and AI in the industry, highlighting their potential impact and future possibilities.

A. AR in Industry

As a rapidly emerging technology, AR is increasingly being adopted across various industrial sectors, providing plenty of potential applications that can improve efficiency, productivity, and safety. There is a considerable amount of AR applications in the industry, and some examples are step-by-step guides [22], manufacturing [4], [23], [24], design [25], [26] and evaluation [27], [28]. Wang et al. [29] highlights in their literature review the need for research in several aspects when applying AR to the industry field, such as knowledge representation and contextual awareness. AR can provide many benefits in product design, allowing for faster and more collaborative tasks [30]–[34]. Process design and control is another field of interest, as indicated by Elia et al. [35], and several applications and systems have been developed [36], [37], bringing to attention the benefits of using AR in this field of application. Regarding maintenance, much interest has been put into solving challenging problems, such as reducing the Mean Time To Repair (MTTR) [38], reducing the cost of having SMEs on site [39], guiding in bad viewing conditions [40], or focusing on the operators' safety [41]. There has also been research about using AR for accessing information, such as the ARES framework, to adapt the information shown to the operator depending on several conditions, such as the time to perform a task [42]. This highlights the importance of the different roles and experiences on the shop floor and how the interface must show more or less information depending on these characteristics. Additionally, the findings indicate that using authoring tools by SMEs makes creating instructions more efficient and user-friendly. For example, Palmarini et al. developed the FARA authoring tool, which facilitates the creation of step-by-step AR animations for various procedures, such as maintenance, repair, and overhaul (MRO) [43].

B. AI in Industry

Equally important, AI is being applied in industrial fields rapidly, alongside AR in various domains. MRO, diagnosis, and predictive maintenance are among the fields where AI has found widespread applications. Predicting a possible error in the system before it happens leverages AI to foresee potential system failures before they occur, thus enabling proactive maintenance and increased operational efficiency. Both Carvalho et al. [18], and Zonta et al. [44] perform a systematic literature review where several ML and DL models are being applied for predictive maintenance, demonstrating the increasing research interest in this field. Other applications focus on customer support, where chatbots and recommendation systems can help companies provide faster and more personalized support to clients. Casillo et al. develop a chatbot framework for real-time assistance and efficient and personalized training [45]. Pattern recognition and prediction are, by nature, key applications of ML and DL algorithms. Detecting patterns in vast amounts of data might help data scientists obtain valuable insights, for example, to predict changes in product demand. Moroff et al. evaluate several ML and DL models such as Random Forest, XGBoost, Long-term short-term memory (LSTM) networks, and a multilayer perceptron (MLP), among others, as forecasting models [46]. Finally, AI models are being increasingly applied in the field of automation. Operators' previous tasks can be automatized intelligently, such as optimizing production processes and improving customer support. Maschler and Weyrich highlight, in their literature review, several studies in fields such as anomaly detection, time series prediction, fault diagnosis and prognostics, quality management, and computer vision [47].

C. Synergy Between AR and AI in Industry

As a complementary element to AR, AI opens new synergy possibilities. In the field of information access, Chidambaram develops a solution utilizing AR and the YOLO foundation model [48] to generate instructions that differentiate between novice users and SMEs [49]. As described by Standford, a foundation model means to *"Train one model on a huge amount of data and adapt it to many applications"*, or in other words, it is a model that has been pre-trained and provides various features that can be utilized for transfer learning or fine-tuning to fit specific requirements [50]. Examples of foundation models are YOLO for object detection, Stable Diffusion for image generation [51] or GPT for text generation [52]. Our previous research has focused on developing tools that guide and enhance the safety of shop-floor operators using AR and AI [12]; however, the present proposal in this work takes a closer look at the other side of the equation, the SMEs and how to use AR and AI to enrich the environment with information in a comfortable manner. Little research has been done regarding the use of these two technologies in enhancing unstructured information management and access, and this work proposes an approach to fill this gap.

D. Documentation Management and Access

With its human-centric view, the advent of Industry 5.0 [53], [54] brings about significant challenges in the realm of documentation and information access, owing to various factors such as decentralization, virtualization, and modularity [8]. This highlights the need for more effective methods for managing documentation. In light of the need for information to be easily accessible, updatable, and translatable, paper-based documentation is becoming obsolete. Further research must be conducted in this area, as several authors have emphasized [14]–[16]. This study seeks to address a key challenge in shop floor operations by exploring novel strategies to enhance information accessibility. The ultimate aim of this proposal is to leverage the knowledge and expertise of SMEs to create dynamic environments, enhancing them with knowledge anchors into spatial 3D real environments to improve efficiency and profitability.

E. Information Retrieval and Mental Decay

In accordance with the discussion presented in section I, the most optimal way to gain expertise in an industrial setting is to perform SBT and personalized tutelage with an SME; however, one of the most critical issues associated with this process is the maintenance of the acquired knowledge, particularly its tendency to deteriorate over time.

Mental decay, also known as knowledge decay, is a passive process in which the knowledge and skills of a person gradually decline over time if not actively reinforced. Studies have shown that mental decay can occur even when an individual is exposed to new information, with decay increasing as the time between exposure and retrieval increases [55]. Numerous studies have been conducted on knowledge retention and information retrieval in industrial settings in recent years. One such study by Adesope et al. found that repeated exposure to information leads to better knowledge retention compared to solitary exposure [56]. This finding is supported by other studies, such as the work of Karpicke and Roediger, who showed that retrieval practice can enhance long-term retention of information [57]. In addition, research has also been conducted on the impact of aging on knowledge retention and retrieval. For example, Bissig and Lustig found that older adults experience greater difficulty retrieving information from long-term memory compared to younger adults [58]. This finding has important implications for industrial settings, as the aging workforce is becoming increasingly prevalent and might be a focus of interest in future research [59].

In industrial environments, knowledge about machines and elements on the shop floor is often distributed through multiple documents and SMEs. Hence, having a reliable source for accessing and retrieving this information is essential. Information access is of paramount importance in industrial settings as it plays a critical role in ensuring the efficient performance of operations. Understanding how the knowledge provided to operators fades over time becomes increasingly important. It is essential to note that the operators involved in the experiments were only given the task to perform with prior knowledge. As explained in section IV, operators were subjected to repeated exposures of the same information because this can lead to better knowledge retention compared to a solitary exposure [60],[61].

The proposed tool aims to fulfill this gap of mental decay, thus providing access to technical and SME information at all times.

III. Proposed System

One of the significant challenges faced by shop floor operators in industrial settings is knowledge retrieval from the environment, as previously discussed in section I and II. It has been discussed that having an SME and utilizing SBT may be the ideal solution, but not always feasible in practice. Furthermore, mental decay adds to the difficulty as the shop floor operator may not always retain all of the information taught. Although AR applications have been proposed as a means of providing additional information in a context-aware system boosted by AI systems; accessing information naturally when technical documentation and SMEs are the only sources of information remains a challenge. This section presents a detailed description of the proposed system, highlighting the key roles of the SME, the shop floor operator, and their interactions with the system. This research aims to address the gap in the literature and justify the main contributions outlined in section I.

A. SME: Context Enrichment With Information

The SME is an expert who has acquired extensive knowledge in a particular field or topic; however, disseminating their knowledge and its contribution to the field remains challenging. While the possibility to ask the SME in case there is doubt exists, it may only sometimes be feasible, as the constant presence of an SME in the work environment may not be practical [11], [39]. Indeed, AR systems can reduce the cost of having SMEs on-site, but the challenge remains in effectively transferring the SME knowledge to the worksite. To bridge the gap in knowledge transfer to the site, this study proposes an architecture that considers the SMEs roles as a "Knowledge Transfer Experts".

The SMEs are responsible for utilizing the system to introduce "pills" of knowledge across the environment. In this research paper, a "pill" refers to a small unit of knowledge that can be added to a system. The term is chosen for its memorable connotation and aligns with the concept of intentional knowledge management. It is important to highlight that the presented architecture implementation relies on the fact that the environment needs to be previously scanned, a common feature in current SLAM-based AR solutions. Upon entering the environment, the SME can interact with their surroundings using touch interaction. This way, the SME can add specific "pills" of information to any element they find interesting to enrich, regardless of whether they are machines, control panels, or any other element of interest in complex environments. This information "pills" will be used by the system with two purposes:

1. To retrieve a specific "pill" linked to a specific position in the environment as-is,
2. For obtaining answers to specific questions.

The present study depicts a specialized tool that supports SMEs in contributing to a digitized environment. The tool facilitates data input through the means of either voice recognition or written text. Using ray-casting techniques, alongside touch interaction in AR, enables SMEs to pinpoint and enrich specific features of the 3D scanned mesh from the virtual environment. The interaction in the system is performed using touch input that is implemented differently depending on the final device used. Specific AR devices can trigger the touch action with hand-specific controls or even by using hands, while for mobile devices like tablets, touching the screen at the desired object triggers the touch action. In both cases, the interaction is implemented using ray-casting, which calculates a line or ray from the touch 2D coordinates and with the direction derived from the camera frustum. Then, the ray intersections are checked, and the object selected is the one that is closest to the user in 3D coordinates. The AR library maintains a congruent mapping of spatial coordinates between the physical and virtual worlds, resulting in accurately identifying elements in the 3D virtual space. A visualization of the SMEs task in the environment is displayed in Fig. 1.

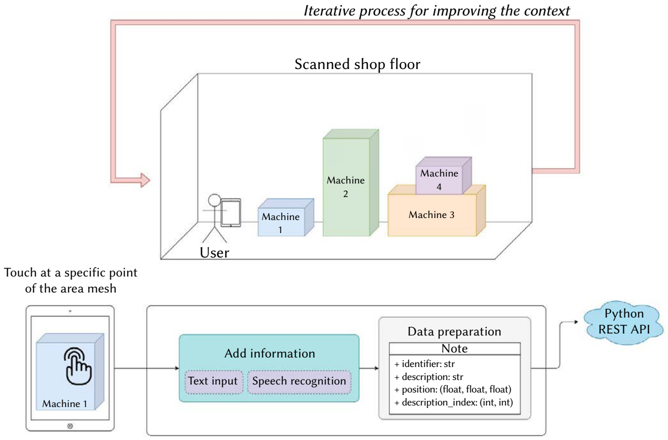


Fig. 1. SMEs annotate scanned environments.

B. Shop Floor Operator: Information Retrieval

Once the site has been enriched with anchored "pills" of knowledge, it is time for the shop floor operator to utilize these resources, reducing the frequency of their need to seek clarification from the SME and technical documentation.

As depicted in Fig. 2, the presented application offers two alternatives for accessing such information in a multimodal manner:

Touch & Area Selection Since many anchors might be disseminated around the site, the shop floor operator has the option to select a rectangular area and retrieve all the "pills" within that selection, as shown in Lane C in Fig. 3. Applying the ray-casting methods as introduced in section III.A, the system can obtain the 3D virtual coordinates of a chosen location on the shop floor by utilizing touch interaction. The process involves the shop floor drawing an area of interest from which anchored "pills" can be retrieved. The approach leverages the same ray-casting techniques employed by SMEs in the aforementioned section, demonstrating the tool's versatility across multiple domains.

Speech Recognition for NL Queries While the *touch and area selection* method is proactive, the system also includes a more reactive approach to obtaining information. The shop floor operator can ask any query in NL, such as "What temperature does glycol evaporate at?". The system will then process the query and provide the answer (e.g., "360º") as well as the location within the work environment where the SME anchored the corresponding "pill" of knowledge. This is also shown in Lane B in Fig. 3.

C. System Implementation

For evaluating the proposed system, a mobile application has been developed using the Unity platform, which allows for developing

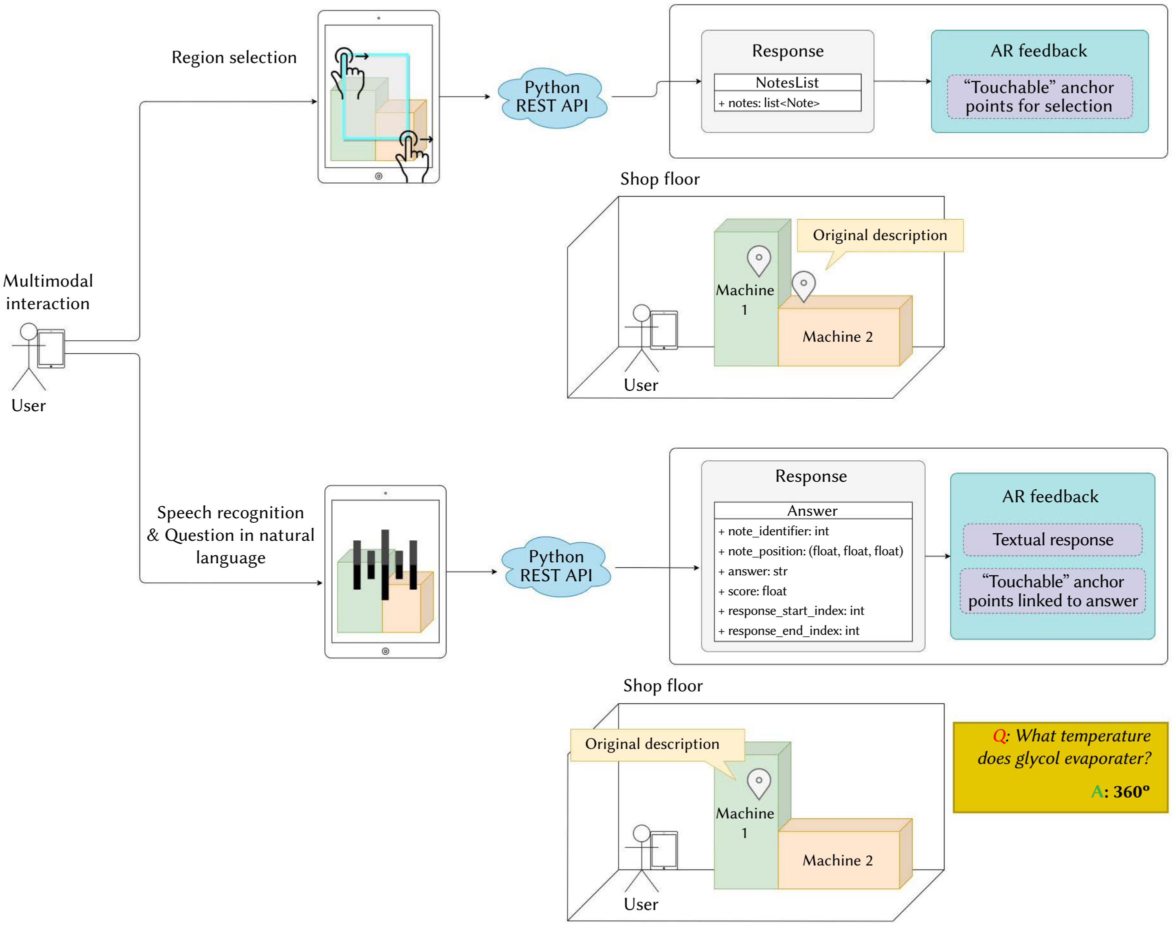


Fig. 2. Shop floor operator retrieves anchored information via NL query or area selection.

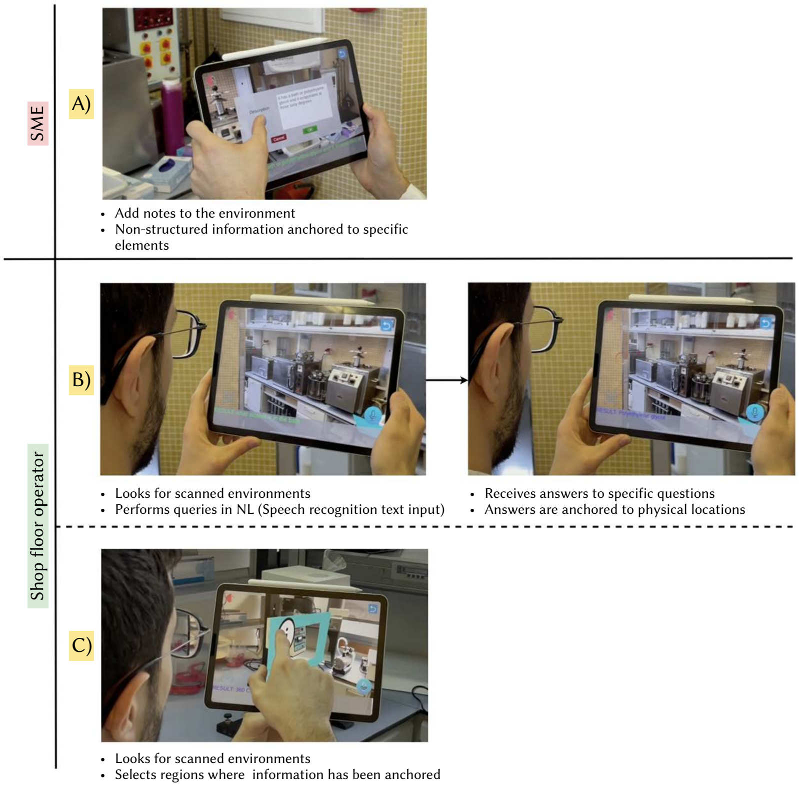

SME
A)
Add notes to the environment
Non-structured information anchored to specific elements
Shop floor operator
B)
Looks for scanned environments
Performs queries in NL (Speech recognition text input)
Receives answers to specific questions
Answers are anchored to physical locations
C)
Looks for scanned environments
Selects regions where information has been anchored

Fig. 3. Two shop floor roles: SMEs add information, operator retrieves it via NL/AR.

applications for both Android and iOS devices. AR through the Vuforia library has been integrated for scene recognition (i.e., the scanned laboratory). For Speech Recognition on device, the Vosk toolkit has been used, which allows for real-time voice recognition in diverse languages. Regarding the server side, since the information must persist between application uses, the Python FastAPI framework has been utilized for generating the different endpoints using a RESTful API.

Fig. 3 briefly summarises the application's key features. Lane A highlights the capabilities of SMEs within the app. After detecting the environment, the SME can add pieces of information, either handwritten or by voice, at any point, using touch interaction. Lanes B and C outline the interaction from the shop floor perspective. In Lane B, the operator can use the application to ask, in NL, about anything regarding contextual information. This query is then sent to the transformer for processing. The specific transformer architecture is explained in detail in subsection D. The result is displayed not only as a text answer but also as a highlighted location in the environment where the SME added the information. Finally, Lane C briefly shows that the operator can retrieve any information in any environment by drawing a rectangle around the area of interest using touch interaction. The application will display all notes anchored by the SME within that area.

D. Consistency of the Information

Ensuring the consistency of textual information units when integrating them into a system is a crucial factor to consider. It is important to estimate whether a new block of textual information is contradictory, redundant, or provides new information content, particularly in a proposal where unstructured information is the fundamental input. The topic of consistency in LLMs is a recurring subject of study, as noted by Elazar et al. in their work on measuring LLMs [62]. In this context, consistency is treated systematically and comprehensively, with a focus on ensuring the correctness of the answers provided by the model prior to paraphrasing the input questions.

In this study, the main focus is on consistency, which involves ensuring that there are no redundancies or contradictions when introducing a new block of information related to a particular aspect of the environment. Redundant information can reduce the efficiency of the model, while contradictory information is even more critical. It is crucial to prevent SMEs from providing conflicting information about the same element, as this can cause problems when correcting subsequent information queries. Such conflicts may arise not only from an SME's error but also from speech-to-text conversion, for instance. To solve this problem, LLMs allow a fine-tuning process after pre-training to improve their behavior when faced with more specific problems. For this specific problem we use a Transformer architecture.

Transformer architecture is a neural network model composed of two key components: an encoder and a decoder. The encoder comprises multiple layers comprising two sub-layers: an attention mechanism and a fully connected feed-forward neural network. The decoder, on the other hand, follows a similar structure. However, in this case, each sub-layer comprises three sub-layers: the attention sub-layer, the feed-forward one, and a third sub-layer in which multi-head attention is applied to the output of the encoder. Using this mechanism makes it possible to have better results than with recurrent networks since it is possible to take into account the semantics of the input sentence more efficiently. In addition, the training process can be unsupervised; however, it is necessary to use significant amounts of unlabelled text. Due to the cost of training a transformer from scratch, not only in terms of time but also in terms of computational units and the amount of data needed to obtain a good performance, it is common to use pre-trained models. That is, using architectures that have already been trained with vast amounts of data. This allows the pre-trained transformers to have already learned most of the semantics of NL, so they can process and answer most of the questions or suggestions that the user asks in NL in a very flexible way. However, it is important to carefully select the dataset to fine-tune the model, as it is of balanced and representative data.

It is possible to find a wide variety of transformers with a wide range of capabilities in NL processing. Such as machine translation between language pairs, text summarisation, text-relevant QA, or conversational systems. Among the most commonly used transformer-type pre-training systems currently in use are DistilBERT [63], RoBERTa [64], Google's T5 [65], BLOOM [66], or GPT-3 [52]. Arroni et al. [67] provide a compelling example of the effective use of LLMs in their work on semantic classification.

In the case of this project, the GPT-JT model was used as resulting of the fine-tuning of the GPT-J model with UL2 training objectives [68], [69], achieving results similar to models of 175B parameters in many tasks, such as InstructGPT davinci v2, but with only 6B parameters [70]. GPT-JT has been trained in a decentralized way and allows its free download and use, as well as its installation and local use. The final LLM used in the final solution can be adapted to specific final requirements of the solution and available options. GPT-JT was a good commitment in evaluating the proposed solution, allowing good results in QA on technical documentation and a high degree of flexibility on few-short learning. Few-shot learning means that the model can be presented with one or several examples of the task to be solved to achieve higher degrees of accuracy in its responses.

With this purpose, the GPT-JT model has been tested to detect if a new piece of textual information is redundant or in contradiction with the previous information assigned to a specific element of the scene of the AR environment.

Although the main aim of this research was not to compare various models on the same instruct-based tasks, we set specific criteria for selecting the most appropriate model. Based on its Open Source availability, superior performance against instruct-based queries (as per the Hugging Face ranking), and ease of installation on local servers (6B parameters only), the GPT-JT model was chosen. The preliminary findings indicated that the chosen model classified the information consistently.

Table I shows a concrete example used to discriminate between "new", "contradictory", or "redundant" with few-shot learning, in this case, a single training example. Using only one example, it has been possible to verify the correctness of the classification of the new information block in all the tested examples. An exhaustive evaluation of this approach transcends the objectives of the present proposal, more typical of computational linguistics, requiring the creation of comprehensive and specific evaluation datasets which, in the best of cases, cannot guarantee their results in the face of domain problems.

The proposal in this paper focuses on using the few-shot learning consistency test of the transformer to detect redundancy or contradiction problems better and visually notify the SME of this possibility. When the SME is warned of a possible redundancy or contradiction, it can examine the text entered associated with an element and repeat in case of a possible error or reconfirm the correction of the new information element.

IV. Evaluation and Results

A. Experimental Setup

Experiments were conducted in a textile laboratory at Universitat Politècnica de València, Campus d'Alcoi. The section of the laboratory that was scanned for subsequent identification is a representative

TABLE I. GPT-JT Tests Label Information as "New", "Redundant", or "Contradictory" Based on Context

Context	Input	Output
Few-shot learning		
To turn on the machine switch on the red button. *To turn off the machine switch off the red button.*	*To turn on the machine it is necessary to switch on the red button.*	"redundant"
Same context.	*To turn on the machine it is necessary to switch on the blue button.*	"contradictory"
Same context.	*To pause the machine it is necessary to switch on the blue button.*	"new"
After few-shot learning...		
The emulsion is homogenized with an agitator. *One field of use is microcapsule emulsions or cosmetic creams.* *At the top of the panel is the button to raise and lower the agitator.* *In the central part of the panel are the buttons to turn on and off, and a wheel to control the number of Revolutions Per Minute (RPM).* *At the bottom, we find the motor and ignition indicators.*	*To lower the agitator, use the button on the top of the panel.*	"redundant"
Same context.	*The machine is called homogenizer.*	"contradictory"
Same context.	*We can find the buttons to turn off the machine at the bottom.*	"new"

sample of an overall facility. It was selected because it contains a variety of equipment commonly used in textile manufacturing and provides a suitable environment for testing the developed system. The equipment used in the experiments includes machines for fabric dye testing, material cleaning, and emulsion homogenization, all essential for producing high-quality textile products. The laboratory's wide range of equipment and diverse capabilities make it an ideal environment for simulating potential scenarios, providing a representative setting for testing and evaluating various approaches and solutions. To evaluate the developed application, we selected an iPad Air device because of its compatibility with the system and development tools and its ease of use for the operators.

B. Participants

As far as participants are concerned, a textile master teacher served as the SME for explaining and adding information. Two groups of participants were selected to evaluate the system. 30 participants were selected and distributed between groups A and B.

The 30 participants recruited for the study were all master's students in engineering, ranging in age from 22 to 28 years old, with an equal distribution of male and female participants. While all participants had previous experience with similar machines, none were familiar with the specific machine used in this study, making it a novel task for all participants.

While both groups were exposed to the explanation of the SME, during the system evaluation phase, group A had access to the documentation about the machines to be utilized and the SME. In contrast, group B had access to the developed application. The only restriction applied to group B was that they were instructed to use the application first for any questions about the machines. If the answer from the application was incorrect or lacked enough information, they were allowed to search in the technical documentation and ask the SME.

Both groups, A and B, were exposed to information about the environment and the machines they were going to use during the experiment. The information exposure took place simultaneously a week before the experiment for both groups. The information was presented by a SME, who, at the same time, introduced the information into the system. Three days before the experiment, both groups were again presented with the same information to boost retention.

C. Tasks

The participants were instructed to perform several interactions with three types of machines; fabric dye testing (Task 1), material cleaning with an ultrasonic machine (Task 2), and homogenizing emulsions (Task 3). Fig. 4 compares the completion time between groups A and B while performing the same tasks in seconds. To ensure objective and consistent measurements, the data was collected by a single external observer who followed standardized procedures throughout the data collection process.

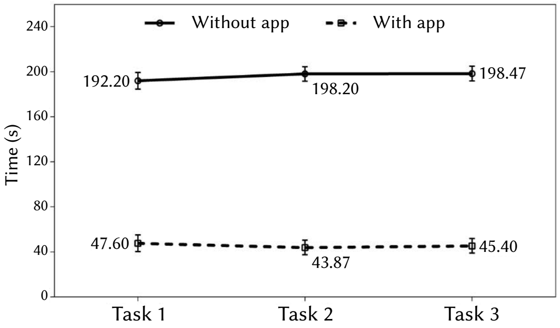


Fig. 4. Time comparison between groups A (No app) and B (With app).

The following standardized procedures were employed by the observer:

1. **Training and Familiarization**: The observer underwent comprehensive training to become familiar with the research goals, the tasks to be performed, and the specific machines involved. This training aimed to ensure that the observer had a thorough understanding of the procedures and requirements for accurate data collection.
2. **Non-Intrusive Observation**: The observer adopted a non-intrusive approach to minimize any potential influence on the participants' behavior or performance. The observer focused on discreetly observing the participants without interfering with their interactions or affecting the natural flow of the tasks. This approach aimed to ensure that the participants' actions were representative of their usual behavior.
3. **Data Recording**: The observer used the same data collection sheet to record relevant information during the observation process. This included capturing the start and end times of each task, any relevant observations or notes, and any additional contextual information that could be important for later analysis.

This approach allowed for precise time measurement and reduced the potential for biases or errors that could arise from multiple observers or inconsistent methods. It is clear that there is a significant reduction in time when specific information needs to be retrieved, thereby supporting the second and third main contributions of this research: anchoring information in the environment reduces the need

TABLE II. Comparison of Time Between Groups A and B, Along With the Corresponding P-values

	Task			**Effects tests**		
	1	**2**	**3**	**Group**	**Tasks**	**Group*Task**
	Mean	*Mean*	*Mean*	***F*** **(g.l.);**	***F*** **(g.l.);**	***F*** **(g.l.);**
	(Sd)	*(Sd)*	*(Sd)*	***p*****-value** (η^2)	***p*****-value** (η^2)	***p*****-value** (η^2)
Time (seconds)				*F* (1;28) = 1.545,22; *p* <0,001 (0,982	*F* ((2;56) = 0,04; *p* = 0,964 (0,001)	*F* ((2;56) = 0,25; *p* = 0,781 (0,009)
Group A	192, 20 (34, 32)	198, 20 (31, 26)	198, 47 (32, 45)			
Group B	47,60 (20, 28)	43, 87 (15, 24)	45, 40 (13, 62)			

TABLE III. Likert Questionnaire

	Min-Max	**Mean (Sd)**
I found the AR app easy to use for accessing information about the machines.	3-5	4 (0,85)
The information provided by the AR app was accurate and reliable.	2-5	3,27 (0,8)
The AR app helped me perform my tasks more efficiently.	3-5	4,6 (0,63)
The information provided by the AR app was useful and relevant to my needs.	4-5	4,8 (0,41)
Information was provided by the AR app quickly when I requested it.	3-5	4,07 (0,88)
I did not encounter errors or issues while using the AR app.	2-5	3,47 (0,74)
I did not need to refer to technical documentation/technical operator in addition to the AR app to find the information I needed.	3-5	3,93 (0,7)
The information provided by the AR app was helpful in completing my tasks.	4-5	4,73 (0,46)
The AR app made it easier to access information compared to other methods I have used in the past. I would	4-6	4,47 (0,64)
be willing to use the AR app on a regular basis as part of my workday.	4-5	4,33 (0,72)

for consulting technical documentation and SME, and the reduction in time between the emergence of a doubt and receiving a response.

D. Results

Table II shows the results of the two-factor ANOVA with repeated measures on one of them performed to determine if the effect of the group influences the task execution time. The results show that there is a statistically significant difference between the groups, regardless of the task (p < 0,001) such that the task execution time for workers using the app (45,62 seconds) was significantly lower than for workers not using the app (196,29 seconds). No differences were observed between tasks (p = 0.964), and the group and task interaction were not significant (p = 0,781). The term "group and task interaction" refers to the relationship between the different groups of participants (Groups A and B) and the specific tasks they performed. In this context, a non-significant interaction suggests that the effect of the group on task execution time was consistent across all tasks. In other words, the app usage had a similar effect regardless of the task performed.

Additionally, a Likert questionnaire of 5 points was delivered to the participants of group B (see table III) to measure their perceptions towards the combination of AR and anchored information retrieval in a multimodal manner and the NL interaction complement in the AR system. The questionnaire had ten questions, with values that ranged between 1 (Strongly disagree) and 5 (Strongly agree). The results support the notion that the participants had a favorable view towards integrating anchored information and its retrieval by AR systems. This indicates that this approach could potentially lead to improved outcomes in future studies and practical applications.

V. Conclusions

The roles of SMEs and shop floor operators are essential in Industry 4.0, but even more in the future advent of Industry 5.0. AR and AI techniques are being applied to improve the efficiency and effectiveness of these roles. However, while the use of AR and AI techniques is receiving much attention, only some studies have investigated the value of SMEs as a source of information. The unstructured nature of this information makes it challenging to manage and integrate with technical documentation. In this study, we developed and evaluated a system to extract information from SMEs that can be integrated with technical documentation. We used state-of-the-art AI architectures such as Transformers and LLMs to perform useful tasks such as QA and multimodal interaction on AR systems. Our results demonstrate the potential of integrating SME information with technical documentation to reduce the time it takes for operators to access relevant information. It is worth noting that Industry 5.0 is a human-centric approach to the industry that emphasizes the value added by people in the manufacturing process. While Industry 4.0 focuses on using advanced technology to automate and optimize production, Industry 5.0 recognizes the importance of operator comfort and satisfaction in the workplace. This approach considers the physical and emotional well-being of the workforce, as well as their creativity, problem-solving abilities, and interpersonal skills. By combining the strengths of both human workers and technology, Industry 5.0 aims to create a harmonious and efficient work environment that benefits all stakeholders.

This study highlights the importance of SMEs' knowledge for improving shop floor operations; however, there is still room for improvement in automating the extraction process and maintaining the accuracy and relevance of the information. Future research could focus on developing more advanced NLP techniques to better extract and organize SMEs' knowledge while ensuring that the information remains up-to-date and reliable.

Although the Vosk toolkit was used for the system implementation for speech recognition, in the presence of noisy or industrial environments, speech-to-text accuracy can be significantly improved by employing the latest Open Source models, such as Whisper [71].

Another area for future research is integrating SME knowledge with technical documentation. It would be beneficial to investigate how different types of information can be presented in a way that is easy to access and use for operators. Additionally, there is potential for integrating AR and AI techniques with SME knowledge to further enhance the efficiency and effectiveness of shop floor operations. For example, by implementing automatic information retrieval methods using object detection models, thus allowing operators to access relevant information while exploring the shop floor quickly.

As Industry 5.0 emphasizes the human-centric nature of manufacturing, it is essential to explore ways to improve operator comfort and satisfaction in the workplace. Future studies could investigate using AR and VR technologies to create more engaging and interactive training materials or wearable technologies to monitor and improve operator well-being.

Appendix

This appendix illustrates the different requests the application can make to the server, shown in Fig. 5. It outlines the possible requests the SME and the shop floor operator can make to the server.

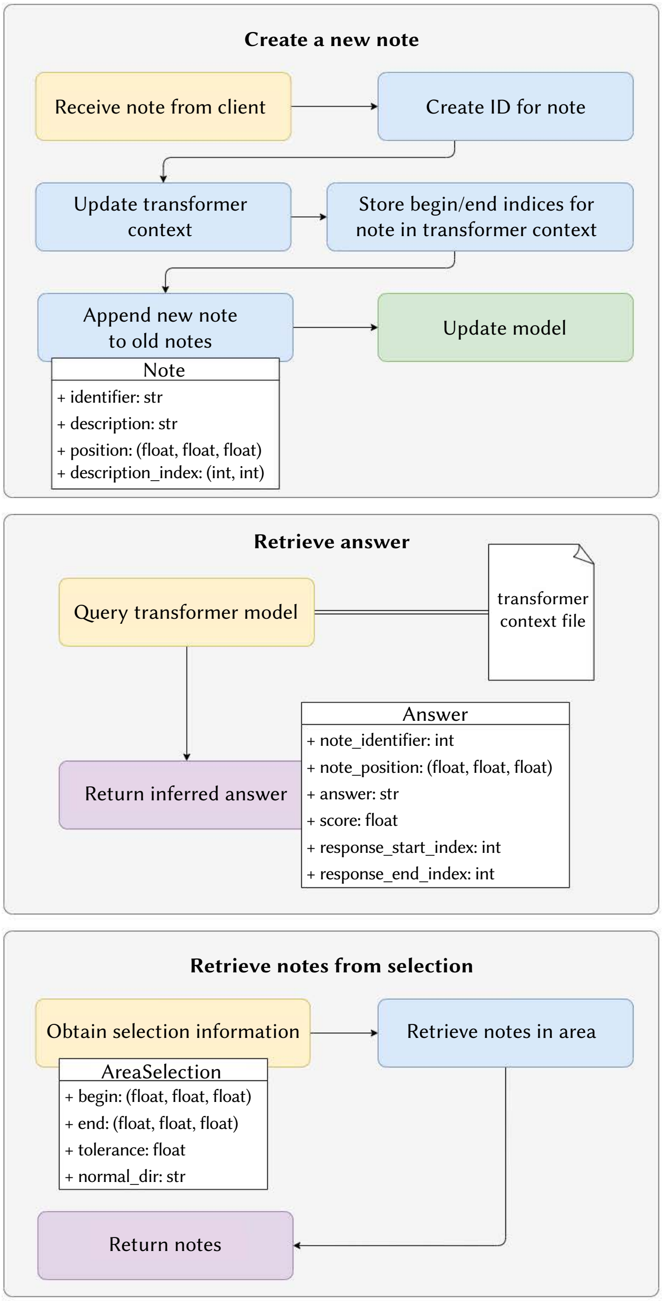

Create a new note
Receive note from client
Create ID for note
Update transformer context
Store begin/end indices for note in transformer context
Append new note to old notes
Update model
Note
+ identifier: str
+ description: str
+ position: (float, float, float)
+ description_index: (int, int)
Retrieve answer
Query transformer model
transformer context file
Return inferred answer
Answer
+ note_identifier: int
+ note_position: (float, float, float)
+ answer: str
+ score: float
+ response_start_index: int
+ response_end_index: int
Retrieve notes from selection
Obtain selection information
Retrieve notes in area
AreaSelection
+ begin: (float, float, float)
+ end: (float, float, float)
+ tolerance: float
+ normal_dir: str
Return notes

Fig. 5. Rest API calls.

References

[1] H. Kagermann, J. Helbig, A. Hellinger, W. Wahlster, *Recommendations for implementing the strategic initiative INDUSTRIE 4.0: Securing the future of German manufacturing industry; final report of the Industrie 4.0 Working Group*. Forschungsunion, 2013.

[2] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, B. MacIntyre, "Recent advances in augmented reality," *IEEE Computer Graphics and Applications*, vol. 21, no. 6, pp. 34–47, 2001, doi: 10.1109/38.963459.

[3] M. Billinghurst, A. Clark, G. Lee, "A survey of augmented reality," *Foundations and Trends in Human- Computer Interaction*, vol. 8, no. 2-3, pp. 73–272, 2015, doi: 10.1561/1100000049.

[4] T. Caudell, D. Mizell, "Augmented reality: an application of heads-up display technology to manual manufacturing processes," in *Proceedings of the Twenty- Fifth Hawaii International Conference on System Sciences*, 1992, pp. 659–669. doi: 10.1109/HICSS.1992.183317.

[5] C. H. Chu, L. Wang, S. Liu, Y. Zhang, M. Menozzi, "Augmented reality in smart manufacturing: Enabling collaboration between humans and artificial intelligence," *Journal of Manufacturing Systems*, vol. 61, pp. 658–659, 10 2021, doi: 10.1016/j.jmsy.2021.05.006.

[6] S. Jaschke, "Mobile learning applications for technical vocational and engineering education: The use of competence snippets in laboratory courses and industry 4.0," in *Proceedings of 2014 International Conference on Interactive Collaborative Learning, ICL 2014*, 1 2014, pp. 605–608, Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/ICL.2014.7017840.

[7] E. Marino, L. Barbieri, B. Colacino, A. K. Fleri, F. Bruno, "An Augmented Reality inspection tool to support workers in Industry 4.0 environments," *Computers in Industry*, vol. 127, 5 2021, doi: 10.1016/j.compind.2021.103412.

[8] M. Gattullo, G. W. Scurati, M. Fiorentino, A. E. Uva, F. Ferrise, M. Bordegoni, "Towards augmented reality manuals for industry 4.0: A methodology," *Robotics and Computer-Integrated Manufacturing*, vol. 56, no. March 2018, pp. 276–286, 2019, doi: 10.1016/j.rcim.2018.10.001.

[9] T. Masood, J. Egger, "Augmented reality in support of Industry 4.0—Implementation challenges and success factors," *Robotics and Computer-Integrated Manufacturing*, vol. 58, pp. 181–195, 8 2019, doi: 10.1016/j.rcim.2019.02.003.

[10] X. Xu, Y. Lu, B. Vogel-Heuser, L. Wang, "Industry 4.0 and Industry 5.0—Inception, conception and perception," *Journal of Manufacturing Systems*, vol. 61, pp. 530–535, 10 2021, doi: 10.1016/j.jmsy.2021.10.006.

[11] L. E. Garza, G. Pantoja, P. Ramírez, H. Ramírez, N. Rodríguez, E. González, R. Quintal, J. A. Pérez, "Augmented reality application for the maintenance of a flapper valve of a fuller-kynion type m pump," *Procedia Computer Science*, vol. 25, pp. 154–160, 2013, doi: 10.1016/j.procs.2013.11.019.

[12] J. Izquierdo-Domenech, J. Linares-Pellicer, J. Orta- Lopez, "Towards achieving a high degree of situational awareness and multimodal interaction with AR and semantic AI in industrial applications," *Multimedia Tools and Applications*, vol. 82, pp. 15875–15901, 2023, doi: 10.1007/s11042-022- 13803-1.

[13] J. Lave, E. Wenger, *Situated learning: Legitimate peripheral participation*. Cambridge university press, 1991. doi: 10.1017/CBO9780511815355.

[14] C. A. Ventura, "Why switch from paper to electronic manuals?," in *Proceedings of the ACM conference on Document processing systems*, 2000, pp. 111–116. doi: 10.1145/62506.62525.

[15] F. Quint, F. Loch, "Using smart glasses to document maintenance processes," *Mensch und Computer 2015–Workshopband*, pp. 203–208, 2015, doi: 10.1515/9783110443905-030.

[16] C. Kollatsch, P. Klimant, "Efficient integration process of production data into Augmented Reality based maintenance of machine tools," *Production Engineering*, vol. 15, pp. 311–319, 6 2021, doi: 10.1007/s11740-021-01026-6.

[17] M. Hermann, T. Pentek, B. Otto, "Design principles for industrie 4.0 scenarios," in *Proceedings of the Annual Hawaii International Conference on System Sciences*, vol. 2016-March, 3 2016, pp. 3928–3937, IEEE Computer Society. doi: 10.1109/HICSS.2016.488.

[18] T. P. Carvalho, F. A. Soares, R. Vita, R. d. P. Francisco, J. P. Basto, S. G. Alcalá, "A systematic literature review of machine learning methods applied to predictive maintenance," *Computers and Industrial Engineering*, vol. 137, 11 2019, doi: 10.1016/j.cie.2019.106024.

[19] J. Ribeiro, R. Lima, T. Eckhardt, S. Paiva, "Robotic Process Automation and Artificial Intelligence in J. Sääski, T. Salonen, M. Hakkarainen, S.

Siltanen, C. Woodward, J. Lempiäinen, "Integration of design and assembly using augmented reality," in *Micro- Assembly Technologies and Applications: IFIP TC5 WG5. 5 Fourth International Precision Assembly Seminar (IPAS'2008)*, Chamonix, France, 2008, pp. 295–404, Springer. doi: 10.1007/978-0-387-77405-3_39.

[20] T. Salonen, J. Sääski, C. Woodward, O. Korkalo, I. Marstio, K. Rainio, "Data pipeline from CAD to AR based assembly instructions," in *Proceedings of the ASME/AFM World Conference on Innovative Virtual Reality 2009, WINVR2009*, 2009, pp. 165–168. doi: 10.1115/WINVR2009-705.

[21] M. Fiorentino, G. Monno, A. E. Uva, "Tangible digital master for product lifecycle management in augmented reality," *International Journal on Interactive Design and Manufacturing*, vol. 3, no. 2, pp. 121–129, 2009, doi: 10.1007/s12008-009-0062-z.

[22] M. Fiorentino, R. Radkowski, C. Stritzke, A. E. Uva, G. Monno, "Design review of CAD assemblies using bimanual natural interface," *International Journal on Interactive Design and Manufacturing*, vol. 7, pp. 249– 260, 11 2013, doi: 10.1007/s12008-012-0179-3.

[23] L. Hou, X. Wang, L. Bernold, P. E. D. Love, "Using Animated Augmented Reality to Cognitively Guide Assembly," *Journal of Computing in Civil Engineering*, vol. 27, pp. 439–451, 9 2013, doi: 10.1061/(asce)cp.1943-5487.0000184.

[24] L. Hou, X. Wang, M. Truijens, "Using Augmented Reality to Facilitate Piping Assembly: An Experiment-Based Evaluation," *Journal of Computing in Civil Engineering*, vol. 29, 1 2015, doi: 10.1061/(ASCE)CP.1943-5487.0000344.

[25] X. Wang, S. K. Ong, A. Y. Nee, "A comprehensive survey of augmented reality assembly research," *Advances in Manufacturing*, vol. 4, pp. 1–22, 3 2016, doi: 10.1007/s40436-015-0131-4.

[26] Industry 4.0 - A Literature review," in *Procedia Computer Science*, vol. 181, 2021, pp. 51–58, Elsevier B.V. doi: 10.1016/j.procs.2021.01.104.

[27] A. Bécue, I. Praça, J. Gama, "Artificial intelligence, cyber-threats and Industry 4.0: challenges and opportunities," *Artificial Intelligence Review*, vol. 54, pp. 3849–3886, 6 2021, doi: 10.1007/s10462-020-09942-2.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention Is All You Need," in *Advances in neural information processing systems*, vol. 30, 2017, pp. 6000– 6010. doi: 10.48550/arXiv.1706.03762.

[29] G. W. Scurati, M. Gattullo, M. Fiorentino, F. Ferrise, M. Bordegoni, A. E. Uva, "Converting maintenance actions into standard symbols for Augmented Reality applications in Industry 4.0," *Computers in Industry*, vol. 98, pp. 68–79, 6 2018, doi: 10.1016/j.compind.2018.02.001.

[30] D. K. Baroroh, C. H. Chu, L. Wang, "Systematic literature review on augmented reality in smart manufacturing: Collaboration between human and computational intelligence," *Journal of Manufacturing Systems*, vol. 61, pp. 696–711, 10 2021, doi: 10.1016/j.jmsy.2020.10.017.

[31] P. Wang, X. Bai, M. Billinghurst, S. Zhang, X. Zhang, S. Wang, W. He, Y. Yan, H. Ji, "AR/MR Remote Collaboration on Physical Tasks: A Review," *Robotics and Computer-Integrated Manufacturing*, vol. 72, 12 2021, doi: 10.1016/j.rcim.2020.102071.

[32] M. Sereno, X. Wang, L. Besancon, M. J. McGuffin, T. Isenberg, "Collaborative Work in Augmented Reality: A Survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, pp. 2530–2549, 6 2022, doi: 10.1109/TVCG.2020.3032761.

[33] B. Marques, S. Silva, J. Alves, T. Araujo, P. Dias, B. S. Santos, "A Conceptual Model and Taxonomy for Collaborative Augmented Reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, pp. 5113–5133, 12 2022, doi: 10.1109/TVCG.2021.3101545.

[34] B. Marques, S. Silva, J. Alves, A. Rocha, P. Dias, B. S. Santos, "Remote collaboration in maintenance contexts using augmented reality: insights from a participatory process," *International Journal on Interactive Design and Manufacturing*, vol. 16, pp. 419–438, 3 2022, doi: 10.1007/s12008-021-00798-6.

[35] V. Elia, M. G. Gnoni, A. Lanzilotto, "Evaluating the application of augmented reality devices in manufacturing from a process point of view: An AHP based model," *Expert Systems with Applications*, vol. 63, pp. 187–197, 11 2016, doi: 10.1016/j.eswa.2016.07.006.

[36] M. L. Yuan, S. K. Ong, A. Y. Nee, "Augmented reality for assembly guidance using a virtual interactive tool," *International Journal of Production Research*, vol. 46, pp. 1745–1767, 4 2008, doi: 10.1080/00207540600972935.

[37] S. K. Ong, Z. B. Wang, "Augmented assembly technologies based on 3D bare-hand interaction," *CIRP Annals - Manufacturing Technology*, vol. 60, no. 1, pp. 1– 4, 2011, doi: 10.1016/j.cirp.2011.03.001.

[38] D. Mourtzis, V. Siatras, J. Angelopoulos, "Real-time remote maintenance support based on augmented reality (AR)," *Applied Sciences (Switzerland)*, vol. 10, 3 2020, doi: 10.3390/app10051855.

[39] A. Gilchrist, "Introducing Industry 4.0," in *Industry 4.0*, Springer, 2016, ch. 13, pp. 195–215, doi: 10.1007/978-1-4842-2047-4.

[40] Z. Ziaei, A. Hahto, J. Mattila, M. Siuko, L. Semeraro, "Real-time markerless Augmented Reality for Remote Handling system in bad viewing conditions," *Fusion Engineering and Design*, vol. 86, pp. 2033–2038, 10 2011, doi: 10.1016/j.fusengdes.2010.12.082.

[41] D. Tatić, B. Tešić, "The application of augmented reality technologies for the improvement of occupational safety in an industrial environment," *Computers in Industry*, vol. 85, pp. 1–10, 2 2017, doi: 10.1016/j.compind.2016.11.004.

[42] A. Syberfeldt, O. Danielsson, M. Holm, L. Wang, "Dynamic Operator Instructions Based on Augmented Reality and Rule-based Expert Systems," *Procedia CIRP*, vol. 41, pp. 346–351, 2016, doi: 10.1016/j.procir.2015.12.113.

[43] R. Palmarini, I. Fernández, D. Amo, D. Ariansyah, S. Khan, J. A. Erkoyuncu, R. Roy, "Fast Augmented Reality Authoring: Fast Creation of AR step-by-step Procedures for Maintenance Operations," *IEEE Access*, vol. 11, pp. 8407-8421, 2023, doi: 10.1109/ACCESS.2023.3235871.

[44] T. Zonta, C. A. da Costa, R. da Rosa Righi, M. J. de Lima, E. S. da Trindade, G. P. Li, "Predictive maintenance in the Industry 4.0: A systematic literature review," *Computers and Industrial Engineering*, vol. 150, 12 2020, doi: 10.1016/j.cie.2020.106889.

[45] M. Casillo, F. Colace, L. Fabbri, M. Lombardi, A. Romano, D. Santaniello, "Chatbot in industry 4.0: An approach for training new employees," in *Proceedings of 2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering, TALE 2020*, 12 2020, pp. 371–376, Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/TALE48869.2020.9368339.

[46] N. U. Moroff, E. Kurt, J. Kamphues, "Machine Learning and Statistics: A Study for assessing innovative Demand Forecasting Models," *Procedia Computer Science*, vol. 180, pp. 40–49, 2021, doi: 10.1016/j.procs.2021.01.127.

[47] B. Maschler, M. Weyrich, "Deep Transfer Learning for Industrial Automation: A Review and Discussion of New Techniques for Data-Driven Machine Learning," *IEEE Industrial Electronics Magazine*, vol. 15, pp. 65–75, 6 2021, doi: 10.1109/MIE.2020.3034884.

[48] J. Redmon, A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv preprint arXiv:1804.02767*, 2018, doi: 10.48550/arXiv.1804.02767.

[49] S. Chidambaram, H. Huang, F. He, X. Qian, A. M. Villanueva, T. S. Redick, W. Stuerzlinger, K. Ramani, "ProcessAR: An augmented reality-based tool to create in-situ procedural 2D/3D AR Instructions," in *DIS 2021 - Proceedings of the 2021 ACM Designing Interactive Systems Conference: Nowhere and Everywhere*, 6 2021, pp. 234–249, Association for Computing Machinery, Inc. doi: 10.1145/3461778.3462126.

[50] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, "On the Opportunities and Risks of Foundation Models," *arXiv preprint arXiv:2108.07258*, 8 2021, doi: 10.48550/arXiv.2108.07258.

[51] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695. doi: 10.48550/arXiv.2112.10752.

[52] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, "Improving Language Understanding by Generative Pre-Training," *OpenAI*, 2018.

[53] J. M. Rožanec, I. Novalija, P. Zajec, K. Kenda, H. Tavakoli Ghinani, S. Suh, E. Veliou, D. Papamartzivanos, T. Giannetsos, S. A. Menesidou, R. Alonso, N. Cauli, A. Meloni, D. R. Recupero, D. Kyriazis, G. Sofianidis, S. Theodoropoulos, B. Fortuna, D. Mladenić, J. Soldatos, "Human-centric artificial intelligence architecture for industry 5.0 applications," *International Journal of Production Research*, vol. 61, no. 20, pp. 6847–6872, 2022, doi: 10.1080/00207543.2022.2138611.

[54] A. Akundi, D. Euresti, S. Luna, W. Ankobiah, A. Lopes, I. Edinbarough, "State of Industry 5.0—Analysis and Identification of Current Research Trends," *Applied System Innovation*, vol. 5, 2 2022, doi: 10.3390/asi5010027.

[55] O. Hardt, K. Nader, L. Nadel, "Decay happens: The role of active forgetting in memory," *Trends in Cognitive Sciences*, vol. 17, pp. 111–120, 3 2013, doi:

10.1016/j.tics.2013.01.001.

[56] O. O. Adesope, D. A. Trevisan, N. Sundararajan, "Rethinking the Use of Tests: A Meta-Analysis of Practice Testing," *Review of Educational Research*, vol. 87, pp. 659–701, 6 2017, doi: 10.3102/0034654316689306.

[57] J. D. Karpicke, H. L. Roediger, "The critical importance of retrieval for learning," *Science*, vol. 319, pp. 966–968, 2 2008, doi: 10.1126/science.1152408.

[58] D. Bissig, C. Lustig, "Who benefits from memory training?," *Psychological Science*, vol. 18, pp. 720–726, 8 2007, doi: 10.1111/j.1467-9280.2007.01966.x.

[59] M. Wolf, M. Kleindienst, C. Ramsauer, C. Zierler, E. Winter, "Current and future industrial challenges: demographic change and measures for elderly workers in industry 4.0," *Annals of the Faculty of Engineering Hunedoara*, vol. 16, no. 1, pp. 67–76, 2018.

[60] N. J. Cepeda, H. Pashler, E. Vul, J. T. Wixted, D. Rohrer, "Distributed practice in verbal recall tasks: A review and quantitative synthesis," *Psychological Bulletin*, vol. 132, pp. 354–380, 5 2006, doi: 10.1037/0033-2909.132.3.354.

[61] S. K. Carpenter, N. J. Cepeda, D. Rohrer, S. H. Kang, H. Pashler, "Using Spacing to Enhance Diverse Forms of Learning: Review of Recent Research and Implications for Instruction," *Educational Psychology Review*, vol. 24, pp. 369–378, 9 2012, doi: 10.1007/s10648-012-9205-z.

[62] Y. Elazar, N. Kassner, S. Ravfogel, A. Ravichander, E. Hovy, H. Schütze, Y. Goldberg, "Measuring and Improving Consistency in Pretrained Language Models," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1012–1031, 12 2021, doi: 10.1162/tacl_a_00410.

[63] V. Sanh, L. Debut, J. Chaumond, T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 10 2019, doi: 10.48550/arXiv.1910.01108.

[64] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 7 2019, doi: 10.48550/arXiv.1907.11692.

[65] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *The Journal of Machine Learning Research*, vol. 21, pp. 5485–5551, 10 2020, doi: 10.48550/arXiv.1910.10683.

[66] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model," *arXiv preprint arXiv:2211.05100*, 11 2022, doi: 10.48550/arXiv.2211.05100.

[67] S. Arroni, Y. Galán, X. Guzmán-Guzmán, E. R. Nuñez-Valdez, A. Gómez, "Sentiment Analysis and Classification of Hotel Opinions in Twitter With the Transformer Architecture," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 1, p. 53-63, 2023, doi: 10.9781/ijimai.2023.02.005.

[68] Y. Tay, M. Dehghani, V. Q. Tran, X. Garcia, J. Wei, X. Wang, H. W. Chung, D. Bahri, T. Schuster, H. S. Zheng, D. Zhou, N. Houlsby, D. Metzler, "UL2: Unifying Language Learning Paradigms," *arXiv preprint arXiv:2205.05131*, 5 2022, doi: 10.48550/arXiv.2205.05131.

[69] Y. Tay, J. Wei, H. W. Chung, V. Q. Tran, D. R. So, S. Shakeri, X. Garcia, H. S. Zheng, J. Rao, Chowdhery, D. Zhou, D. Metzler, S. Petrov, N. Houlsby, Q. V. Le, M. Dehghani, "Transcending Scaling Laws with 0.1% Extra Compute," *arXiv preprint arXiv:2210.11399*, 10 2022, doi: 10.48550/arXiv.2210.11399.

[70] Together, "GPT-JT," 2022. [Online]. Available: https://huggingface.co/togethercomputer/GPT-JT-6B-v1.

[71] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," *arXiv preprint arXiv:2212.04356*, 12 2022.

Juan Izquierdo-Domenech

Juan Jesús Izquierdo Doménech is an Adjunct Professor of Computer Science in Universitat Politècnica de València. He received his Bachelor's degree in Computer Science Engineering from Universitat Politècnica de València (UPV, Spain) and holds a Master's degree in Multimedia Applications from Universitat Oberta de Catalunya (UOC, Spain). He is currently performing his Ph.D. studies in UPV in the field of Human-Computer Interaction, Mixed Reality, and Artificial Intelligence.

Jordi Linares-Pellicer

Jordi Linares Pellicer is an Associate Professor at Universitat Politècnica de València (UPV, Spain), where he leads the VertexLit research group at the Valencian Research Institute for Artificial Intelligence (VRAIN). He received his Ph.D. in Computer Science from UPV and holds a Master's degree in Artificial Intelligence from Universidad Internacional de La Rioja (UNIR, Spain).

Isabel Ferri-Molla

Isabel Ferri Mollá is currently pursuing her Master's Degree in Artificial Intelligence, Pattern Recognition, and Digital Imaging at Universitat Politècnica de València (UPV, Spain). She received her Bachelor's Degree in Computer Science Engineering from UPV in 2022. Her research interests include areas of artificial intelligence, augmented reality, and human-computer interaction.

Testing Deep Learning Recommender Systems Models on Synthetic GAN-Generated Datasets

Jesús Bobadilla, Abraham Gutiérrez *

Universidad Politécnica de Madrid, Dpto. Sistemas Informáticos, Madrid (Spain)

* Corresponding author: jesus.bobadilla@upm.es (J. Bobadilla), abraham.gutierrez@upm.es (A. Gutiérrez)

Received 31 March 2023 | Accepted 29 September 2023 | Published 25 October 2023

Abstract

The published method Generative Adversarial Networks for Recommender Systems (GANRS) allows generating data sets for collaborative filtering recommendation systems. The GANRS source code is available along with a representative set of generated datasets. We have tested the GANRS method by creating multiple synthetic datasets from three different real datasets taken as a source. Experiments include variations in the number of users in the synthetic datasets, as well as a different number of samples. We have also selected six state-of-the-art collaborative filtering deep learning models to test both their comparative performance and the GANRS method. The results show a consistent behavior of the generated datasets compared to the source ones; particularly, in the obtained values and trends of the precision and recall quality measures. The tested deep learning models have also performed as expected on all synthetic datasets, making it possible to compare the results with those obtained from the real source data. Future work is proposed, including different cold start scenarios, unbalanced data, and demographic fairness.

Keywords

Collaborative Filtering, Deep Learning, GANRS, Generated Datasets, Recommender Systems, Synthetic Datasets.

DOI: 10.9781/ijimai.2023.10.002

I. Introduction

The personalization field in the Artificial Intelligence area is mainly focused on Recommender Systems (RS). Relevant RS are Netflix, TripAdvisor, Spotify, Google Music, TikTok, etc. RS are usually classified according to their filtering approaches, mainly: demographic [1], content-based [2], context-aware [3], social [4], collaborative (CF) [5] and their ensembles [6]. Demographic RS make recommendations based on demographic similarities (gender, age, zip code, etc.); content-based RS recommend items with similar content to the consumed ones (book abstracts, product images, etc.). Context-aware filtering usually uses geographic information, such as GPS coordinates. Social filtering relies on followed, followers, etc. CF uses datasets containing the ratings that each user has voted to each item. Ratings can be explicit votes or implicit interactions (clicks, music listened to, films watched, etc.). Of the existing filtering approaches, CF is the most relevant since it provides the most accurate results. The early approaches to CF used the K-Nearest Neighbors algorithm [7]; it is easy to understand and directly implements the concept of CF, but it is also a slow memory-based method, and its results are not accurate compared to modern model-based approaches. The Matrix Factorization (MF) model [8] creates compressed representations of the input data, called hidden factors, and then combines these latent space vectors using the dot product to obtain each user to item prediction. Probabilistic MF and its variations (NMF [9], BNMF, etc.) provide straightforward models that return accurate prediction and recommendations. Furthermore, once the MF model has been trained, it can make very fast predictions compared to the KNN method.

Currently, deep learning approaches dominate the RS research scenario. The simplest deep learning CF model is the Deep Matrix Factorization (DeepMF) [10], where iterative MF learning is replaced with two different neural embedding layers: one for code users and the other for code items. The embedding layers activation maps play the role of the MF hidden factors, where large, discrete, and sparse input vectors are converted to short, continuous, and dense latent space vectors. As in the MF case, the embedding vectors are combined using a dot layer. The variational design of the DeepMF model is called VDeepMF [11], where a Gaussian stochastic noise is introduced after the embedding layers to obtain more robust results. Neural Collaborative Filtering (NCF) [12] is a reasonable extension of the DeepMF model; NCF replaces the dot layer by a Multi-Layer Perceptron (MLP), providing a deep and non-linear combination of the embedding representations. Both the DeepMF and the NCF models improve the MF results.

RS prediction is a regression task where real values are obtained; however, RS recommendation usually is a classification task, where only a discrete number of fixed values can be returned (e.g. number of stars). Then, deep learning classification approaches naturally fit the CF aims; a classification-based deep learning model [13] is proposed

Please cite this article in press as:
J. Bobadilla, A. Gutiérrez. Testing Deep Learning Recommender Systems Models on Synthetic GAN-Generated Datasets, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 30-39, 2025, http://dx.doi.org/10.9781/ijimai.2023.10.002

TABLE I. Comparison Table of Current RS Methods to Create CF Synthetic Data

	method	parameterization	Accuracy	Performance
GANRS	generative	high	high	high
RecSim	generative	low	high	high
Virtual-Taobao	generative	low	middle	high
DataGenCars	statistical	high	low	high
SynEvaRec	statistical	high	low	low

to both implement the recommendation task and provide a reliability value for each recommended item. Additionally, the regular deep classification approach can be improved by combining the obtained <reliability, rating> tuple values [14].

This paper focuses on testing the Generative Adversarial Networks for Recommender Systems (GANRS) [15] generated datasets by applying a representative set of deep learning CF baselines and comparing their recommendation quality results. Generative adversarial networks (GAN) have recently been introduced in the RS area [16] to reinforce the defense strategies of shilling attacks [17], but particularly to improve results by generating augmented data; fake purchase vectors are generated in CFGAN [18] to reinforce the real purchase data. The Wasserstein CFGAN version is the unified GAN (UGAN), and it manages to minimize the GAN collapse mode. Negative sampling information is incorporated in the input data to IPGAN [19], where two different generative models are used, respectively, for positive and negative samples. Temporal patterns have also been combined with GAN models in RecGAN [20], which uses Recurrent Neural Networks (RNN). The reinforcement learning and GAN models are used to process session information rather than rating matrices in the DCFGAN architecture [21]. Conditional rating generation is proposed in [22] by using a Conditional GAN (CGAN). NCGAN [23] uses a GAN to perform recommendation training and a previous neural network stage to obtain the nonlinear features of the users. Finally, unbalanced data sets are processed using the PacGAN concept in the discriminator and a Wasserstein GAN in the generator [24].

Based on Markov chains and recurrent neural networks, RecSim [25] generates synthetic profiles of users and items; its parameterization is low. The social Taobao web site has been used to provide the Virtual-Taobao [26], improving search in this site; internal distributions are simulated by a GAN. RS synthetic data is created using the Java-based generator DataGenCars [27]; it is based on statistical procedures, allowing a flexible parametrization, but returning low accuracy compared to GAN models. Finally, the SynEvaRec [28] framework makes use of the Synthetic Data Vault (SVD) library for RS datasets generation, based on multivariate distributions using copula functions. The SynEvaRec main drawbacks are its poor accuracy and its low performance in the training stage. Table I summarizes the existing methods.

A. Main Contributions

The objective of this paper is to reinforce the existing tests that have been run on the synthetic datasets generated using the GANRS method. Beyond the existing comparatives between source datasets (Movielens, Netflix, and MyAnimeList) and their synthetic versions, attending to their users, items, and ratings distributions, it is convenient to put into the test the generated datasets on real recommendation scenarios. Some specific and limited prediction and recommendation experiments are provided in the GANRS paper [15], but our research extends them with a comprehensive set of recommendation-based tests, where different deep learning models relevant to the CF are used as baselines and significant recommendation quality measures are processed, and their results are compared.

The paper hypothesis is that the GANRS model can adequately mimic different source CF datasets, such as the Movielens family, MyAnimeList, etc., generating synthetic CF datasets that follow the internal patterns and the probability distributions of the source datasets in the deep learning generative processing. The hypothesis is extended to the different parameterizations the GANRS generative model allows, setting a) the number of fake users, b) the number of fake items, and c) the number of samples. We will put the hypothesis to the test by running different deep learning state of the art CF baselines (NCF, DeepMF, etc.) on several GANRS generated datasets and comparing the obtained recommendation qualities. The GANRS synthetic datasets will contain different number of users, items, and samples. Note that if the hypothesis is fulfilled, the GANRS model can be used as a powerful tool to test current and future CF methods and models on challenging synthetic scenarios where the number of users, items and samples can endlessly grow.

In the rest of the paper, section II explains the different deep learning models used in this research, both to generate the synthetic datasets and to test the behavior of baselines on the generated data. Section III introduces the experiments design, synthetic datasets, and baselines. It also shows the results obtained, their explanations, and the discussion. Section IV highlights the main conclusions of the article and the suggested future work.

II. Models

This research uses many deep learning models, both the GANRS [15] generative framework with which the synthetic datasets have been obtained and the different models used to test the generated datasets. These baseline models are as follows: DeepMF [10], VDeepMF [11], regression NCF [12], classification NCF [13], improved classification NCF [14] and binary regression.

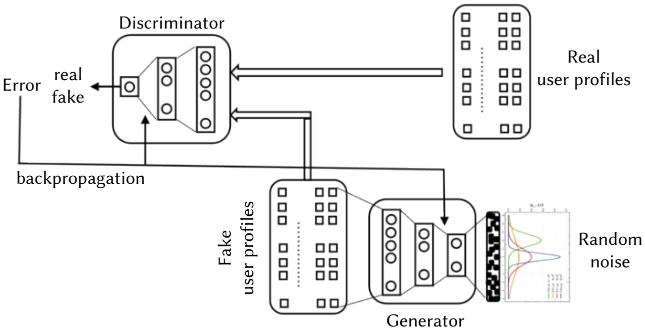


Fig. 1. GANRS architecture.

The GANRS architecture shown in Fig. 1 consists of the generator and discriminator models, where the generator creates CF fake profiles from Gaussian random noise vectors. The discriminator's responsibility is to detect fake samples from training batches of real and fake profiles. Once the RSGAN has been trained from a real source dataset (MovieLens, MyAnimeList, etc.) it can generate as many fake samples as desired by providing the generator with batches of random noise vectors. It is important to note that the GAN is fed with embedded user profiles rather than sparse vectors of ratings. Embeddings are obtained in a previous stage using a DeepMF [10] model.

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

where x are real user profiles and z are random noise vectors (Fig. 1).

$$min_G V(G) = \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \quad (2)$$

$$min_G\ max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \quad (3)$$

The objective of the discriminator can be defined as its ability to recognize real profiles (first term in (1)) combined with its ability to detect fake profiles (second term in (1)). The generator objective is to generate fake profiles that can fool the discriminator (2). Finally, the GAN can be seen as a minimax game in which the discriminator '*D*' tries to maximize *V*, whereas the generator '*G*' tries to minimize it (3).

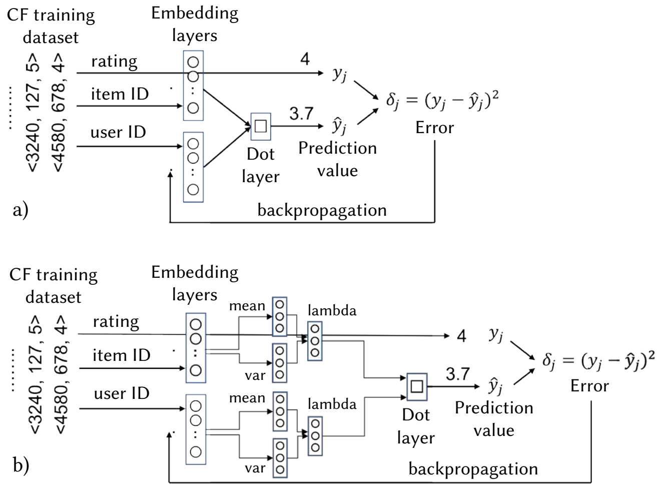


Fig. 2. (a) DeepMF and (b) VDeepMF models.

Regarding the models used to test the generated datasets, the DeepMF and its variational VDeepMF version can be seen as representative baselines. Fig. 2(a) shows the DeepMF model architecture, where two separate embedding layers, one for users and the other for items, convert from discrete and sparse integer inputs to continuous and dense latent space vectors. The hidden factors obtained are combined by means of a dot product layer, as in the MF machine learning model, to predict the rating of each user to each item. The model learns using a loss function that compares each predicted rating with the real label (MSD in Fig. 2).

$$\hat{y}_{ui} = f(P,u,Q,i|P,Q), \text{where P} \in \mathbb{R}^{U\cdot K}, \text{Q} \in \mathbb{R}^{K\cdot N} \quad (4)$$

$$\hat{y}_{ui} = dot(g(u|P), h(i|Q)) \quad (5)$$

$$L_{sqr} = \sum_{(u,i)} (y_{ui} - \hat{y}_{ui})^2 \quad (6)$$

On an RS dataset containing *U* users and *I* items, the prediction of item *i* to user *u* is shown in (4), where the function *f* is defined as a neural network that converts their integer inputs user *u ID* and item *i ID* in their corresponding prediction. *P* and *Q* denote the neural network equivalence to the hidden factors of the MF, where *K* is the number of factors (i.e., the number of neurons in each embedding layer). Note that, usually, the set of weights in *P* and *Q* are called *θ*. The prediction of an item *i* to the user *u* is computed as the dot product of the embedding layer activations g(u|P) and h(i|Q) in (5). Finally, the squared loss is used (6) to learn the model parameters.

The VDeepMF architecture is an extension of the DeepMF one, where a variational stage is added. Fig. 2(b) shows the variational stage located between the embedding layers and the dot layer. This variational stage converts input embeddings to parameters of a statistical distribution (usually a Gaussian one). This concept can be seen in the 'mean' and 'variance' layers that follow the VDeepMF embedding layers, both for users and items (Fig. 2(b)). Each pair of mean and variance layers codes the corresponding Gaussian distribution parameters. Each Lambda layer uses the Gaussian mean and variance to stochastically sample vectors in the latent space. The result is a more robust model due to its stochastic learning.

$$(g(u|P), h(i|Q)) = (v_u, w_i) \longmapsto (\mu_1(v_u), \sigma_1^2(v_u), \mu_2(w_i), \sigma_2^2(w_i)) \in \mathbb{R}^{4K} \quad (7)$$

$$(P_{\mu_1(v_u),\sigma_1^2(v_u)}, Q_{\mu_2(w_i),\sigma_2^2(w_i)}) \quad (8)$$

$$P \sim \mathcal{N}(\mu_1(v_u), diag\ \sigma_1^2(v_u)),\ \ Q \sim \mathcal{N}(\mu_2(w_i), diag\ \sigma_2^2(w_i)) \quad (9)$$

Equation (7) shows the 'mean' and 'var' layers conversion from embedding latent vectors to activation maps representing Gaussian distributions. Thus, the input of the Lambda layers are the pairs of random vectors in equation (8). In equation (9), $\mathcal{N}$ denotes a K-dimensional multivariate distribution, where *μ* represents the mean vector and *diag σ* is the covariance matrix.

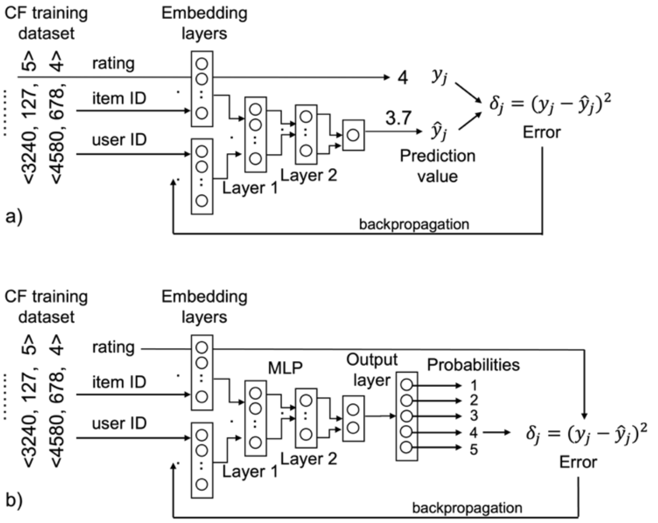


Fig. 3. (a) Regression NCF, and (b) Classification NCF.

The Keras template that summarizes each of the baseline models is provided in Table II. Please note that NCF Binary regression can be coded in a similar way to the regular NCF classification, by replacing the size of the deepest layer to only one output neuron.

The 'regression NCF' term refers to the regular Neural Collaborative Filtering model. This model extends the DeepMF one by adding a Multi-Layer-Perceptron (MLP) stage, as it can be seen in Fig. 3(a). The DeepMF model generates accurate embedding vectors, but it combines them (the user and item vectors) using a linear dot layer. The NCF approach improves the DeepMF model, due to the non-linear and deep learning processing of the embedding output vectors.

$$o_1 = \phi_1(p_u v_u, q_i w_i) \quad (10)$$

$$o_2 = \phi_2(W_2^T o_1 + b_2) \quad (11)$$

$$o_n = \phi_n(W_n^T o_{n-1} + b_{n-1}) \quad (12)$$

$$< \text{r}, \text{v} >= softmax(W_n^T o_n + b_n) \quad (13)$$

$$\hat{y}_{ui} = \text{v} \mid\ < \text{r}, \text{v} > \in\ argmax(\text{r}) \quad (14)$$

$$\hat{y}_{ui} = \sum_{r=1}^{R} r \cdot \text{p} \quad (15)$$

The additional MLP model is formalized in equations (10) to (12), where p_u and q_i denote the embedding layers weights, W_x^T and b_x are the weight matrix and bias vector of layer *x* in the MLP, ϕ_x denotes the layer *x* with its activation function. The regression NCF model has an

TABLE II. KeraS TEMPLATE of the Baseline Models

DeepMF
Input (user) -> Embedding (user) -> Flatten Input (item) -> Embedding (item) -> Flatten Dot (Embedding (user), Embedding (item)) Loss =”mean_squared_error”
VDeepMF
Input (user) -> Embedding (user_mean) -> Dense (user_mean) -> Dense (user_var) -> Lambda($z_{user\_mean} + e^{z\_user\_var} \cdot \epsilon$) -> Flatten Input (item) -> Embedding (item_mean) -> Dense (item_mean) -> Dense (item_var) -> Lambda($z_{item\_mean} + e^{z\_item\_var} \cdot \epsilon$) -> Flatten -> Dot, Loss = “mean_squared_error”
NCF Regression
Input (user) -> Embedding (user) -> Flatten Input (item) -> Embedding (item) -> Flatten -> Concatenate (Embedding (user), Embedding (item)) -> Dense(70) -> Dropout(0.5) -> Dense(30) -> Dropout(0.4) -> Dense(1, “ReLu”) Loss = “mean_squared_error”
NCF Classification
Input (user) -> Embedding (user) -> Flatten Input (item) -> Embedding (item) -> Flatten -> Concatenate (Embedding (user), Embedding (item)) -> Dense(70) -> Dropout(0.5) -> Dense(30) -> Dropout(0.4) -> Dense(6, “softmax”), Loss = “categorical_crossentropy”

output layer containing a unique neuron with an activation function that is linear, implementing the required regression. In contrast, the NCF classification model replaces this output layer with a layer containing as many neurons as possible votes in the RS (usually from one to five stars), as can be seen in Fig. 3(b). The softmax activation function is used in this output layer, while the model loss function is the categorical cross entropy; this ensures a probabilistic output that can be interpreted as a set of <reliability, vote> tuples (13), where the *argmax(reliability)* selects the predicted vote (14). The improved classification model basically combines the existing <reliability, vote> tuple values (15), providing a more accurate output function than the argmax one.

By combining the GANRS generated datasets with the chosen deep learning baselines and the selected recommendation quality measures, a set of experiments is designed and tested in the next section. Results are shown and explained, and finally an overall discussion is provided.

III. Experiments and Results

This paper runs a complete set of experiments to test the performance of current CF deep learning models on GANRS generated datasets.

Table III shows a summary of the designed experiments. The tested CF datasets are generated using ‘GANRS’ [15], obtained from the source datasets: Netflix* [29], MyAnimeList [30], and Movielens 100K [31]. For comparative reasons, results using the three source datasets are also provided. The six deep learning models chosen as baselines are DeepMF [10] and regression NCF [12], and their variations VDeepMF [11], and classification based NCF [13]. Finally, the ‘improved classification NCF’ [14] and the binary regression are included. Since we use classification-based models, where recommendations are not a subset of predictions, only recommendation quality measures can be properly used, from which precision, recall, and F1 have been selected. Finally, we have set even values from 2 to 10 as the number of recommendations (N), and the two most relevant rating values as relevancy threshold (q): 4 & 5 for Movielens and Netflix*, and 9 & 10 for MyAnimeList.

TABLE III. Information Summary of the Designed Experiments

CF deep learning models	CF Datasets	Quality Measures	Testing parameters

DeepMF [10]	Netflix* [29]	Precision	Relevance threshold (q):
			9, 10 (MyAnimeList):
			4, 5 (Netflix* and Movielens).

VDeepMF [11]	GANRS Netflix*: 2,000; 8,000 users	Recall	
Regression NCF [12]	GANRS Netflix*: 150K, 500K, 3M	F1	
Classification NCF [13]	Movielens 100K [31]		Number of recommendations (N): [2, 4, 6, 8, 10]
Classification improved NCF [14]	GANRS Movielens 100K: 2,000; 8,000 users		
Binary regression	MyAnimeList [30]		Gaussian standard deviation: 2.5
	GANRS MyAnimeList: 2,000; 8,000 users		

Table IV shows the values of the main parameters for both the real and synthetic datasets used in the designed experiments. Our first set of experiments are based on the source dataset Netflix*, and it compares the quality recommendation results obtained both from Netflix* and their synthetic generated versions: 2,000 & 8,000 users.

TABLE IV. Main Parameter Values of the Tested Datasets

Dataset	#users	#items	#ratings	scores	sparsity
Movielens 100K	943	1682	99,831	1 to 5	93.71
Netflix*	23,012	1,750	535,421	1 to 5	98.68
MyAnime	19,179	2,692	548,967	1 to 10	98.94
GANRS Netflix* 2,000	2,000	4,000	405,539	1 to 5	94.93
GANRS Netflix* 8,000	8,000	4,000	628,194	1 to 5	98,03
GANRS Netflix* 150K	2,000	4,000	108,710	1 to 5	98,64
GANRS Netflix* 500K	2,000	4,000	272,853	1 to 5	96,59
GANRS Netflix* 3M	2,000	4,000	587,651	1 to 5	92,65
GANRS Movielens 2,000	2,000	4,000	353,269	1 to 5	95,58
GANRS Movielens 8,000	8,000	4,000	509,193	1 to 5	98,40
GANRS MyAnime 2,000	2,000	4,000	419,234	1 to 10	94,76
GANRS MyAnime 8,000	8,000	4,000	654,247	1 to 10	97,95

The three rows in Fig. 4 show, respectively, the results on Netflix* (top row), on GANRS 2,000 users (middle row), and on GANRS 8,000 users (bottom row). The middle and right columns show the precision

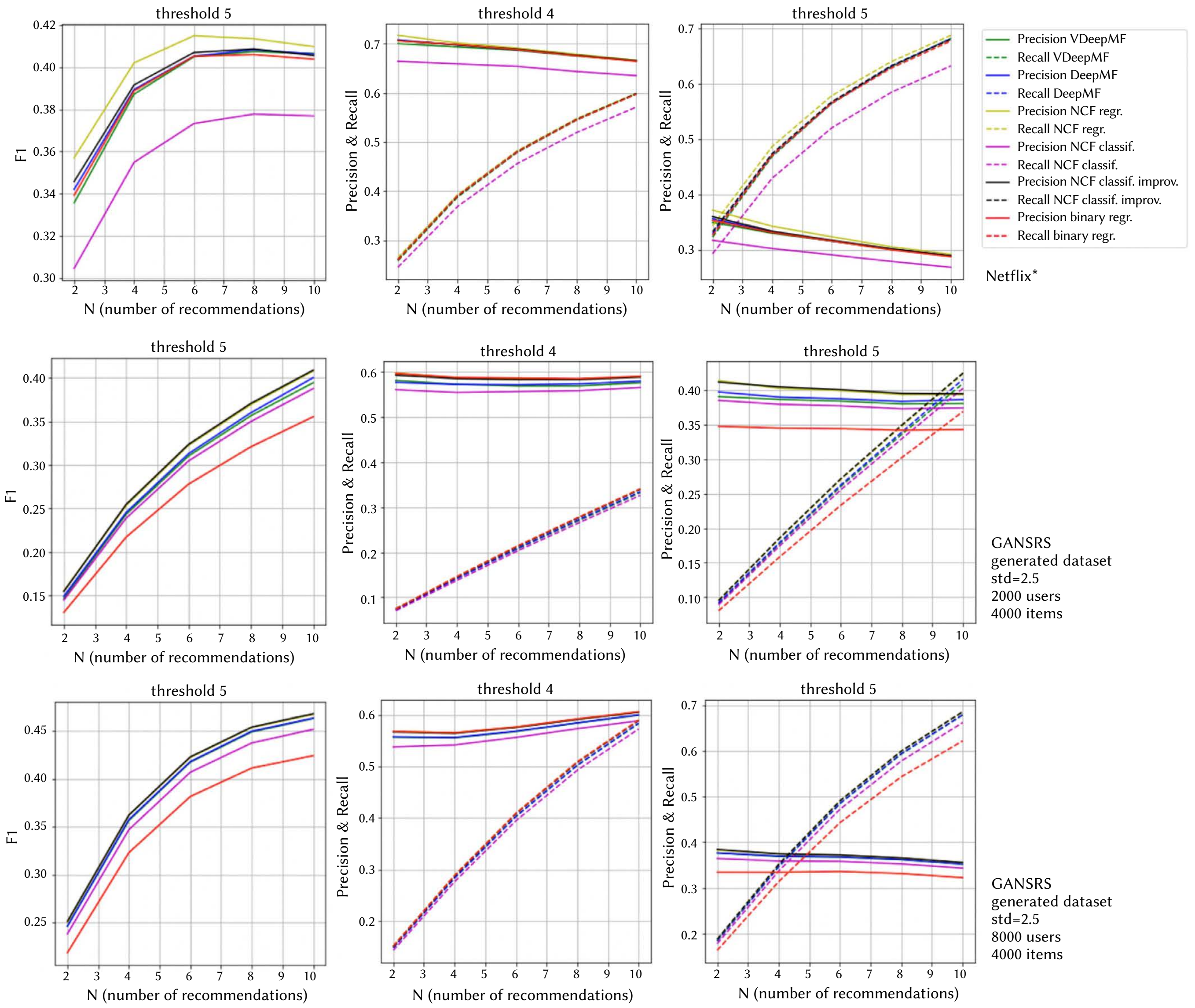


Fig. 4. Comparative among Netflix*, GANRS 2,000 users, and GANRS 8,000 users. Generated datasets include 4000 items and sets 2.5 for the standard deviation of the Gaussian random noise. Number of recommendations N = [2, 4, 6, 8, 10].

and recall values when threshold q is set to 4 and 5 (respectively). The left column shows the precision/recall based F1 quality measure. The legend in the upper-right area of Fig. 4 holds the colors that represent each one of the chosen deep learning baselines. Note that the expected behavior is the superior performance of the deep learning models: regression NCF, improved NCF classification, VDeepMF and DeepMF, whereas classification NCF and binary regression should provide weaker results.

A. Experiment 1: Netflix* Versus GANRS 2000 Users, Versus GANRS 8000 Users

This experiment compares the absolute values and the trends in the recommendation quality obtained for each baseline when applied to the original Netflix* dataset, to the GANRS generated dataset setting 2000 users, and to the GANRS generated dataset setting 8000 users. Both generated datasets take Netflix* as the source to catch its internal patterns. We expect similar trends in the graph functions, showing that the GANRS generated datasets adequately mimic the Netflix* patterns. We also expect different absolute quality values due to the different number of users selected for each GANRS generated dataset.

The top row in Fig. 4 (Netflix*) shows the expected performance evolutions, where the higher the number of recommendations (x-axis), the lower the prediction quality measure, and the higher the recall (it is more complicated to get right 10 recommendations than to get right the two most promising ones). In the same way, a lower threshold value (middle graph) gets a better precision than a higher threshold value (right graph), since there are more samples that reach the threshold, and consequently it is easier to get right with the recommended items. In contrast, the higher the threshold, the better the recall, since there will be less relevant items in the recall denominator. Once we have checked the expected behaviors, the key question is: will the synthetic datasets accomplish the expected trends? Looking at the middle and bottom rows in Fig. 4 we can observe the same aforementioned tendency. The relevant difference between the results from the source Netflix* and the generated GANRS is not the quality trend, but the absolute precision and recall values, where the precision is slightly superior in the GANRS datasets, whereas recall is lower. Please note that the Netflix* dataset contains 23,012 users (Table IV), and then the GANRS versions, particularly the 2000 user version, suffer from a lack of richness that influences the recall results. Additionally, as expected, the higher the threshold, the worse the precision and the better the recall.

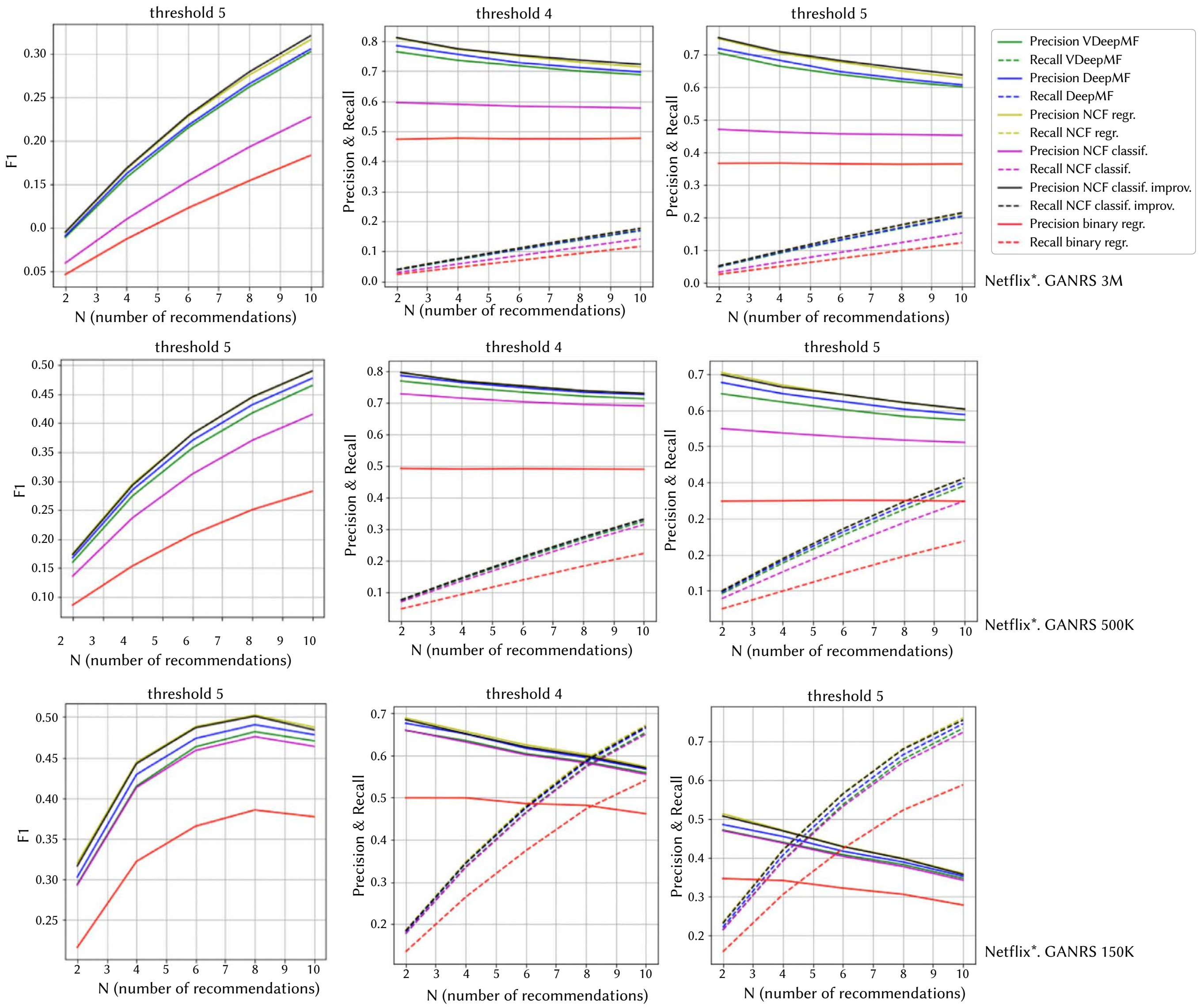


Fig. 5. Comparative among Netflix*, GANRS 150K, 500K, and 3M samples. Generated datasets with 2,000 users; 4,000 items and 2.5 for the standard deviation of the Gaussian random noise. Number of recommendations N = [2, 4, 6, 8, 10].

The F1 quality measure (left column in Fig. 4) balances precision and recall and allows us to visually compare the different results of the data set. We can observe that synthetic datasets provide quality trends and values that are compatible with those achieved by their source dataset (Netflix*). In addition, the GANRS 8,000 user results are more similar to the source than the GANRS 2,000 user ones, as expected due to the 23,000 users contained in Netflix*. Regarding the behavior of deep learning baselines, synthetic data sets maintain the 'ranking order' obtained from Netflix*, where the regression NCF slightly 'wins', closely followed by the improved classification NCF, DeepMF and VDeepMF. NCF classification and binary regression swap their position in the queue when tested on Netflix* and GANRS. Overall, synthetic GANRS datasets perform adequately for CF testing using state-of-the-art deep learning models.

B. Experiment 2: Netflix Based GANRS 3 Million Samples, Versus GANRS 500 Thousand Samples, Versus GANRS 150 Thousand Samples*

This experiment compares the absolute values and the trends in the recommendation quality obtained for each baseline when applied to the GANRS generated dataset setting 3 million samples, to the GANRS generated dataset setting 500 thousand samples, and to the GANRS generated dataset setting 150 thousand samples. All the generated datasets take Netflix* as the source to catch its internal patterns. Like the previous experiment, we expect similar trends in the graph functions, showing that the GANRS generated datasets adequately mimic the Netflix* patterns. We also expect different absolute quality values due to the different number of samples selected for each GANRS generated dataset.

The following experiment uses three synthetic GANRS datasets where the number of samples varies. We use the GANRS Netflix* 150K, 500K, and 3M versions (Table IV) which, respectively, contain 108710, 272853 and 587651 samples. Fig. 5 shows the recommendation results obtained in the 3M version (top row), the 500K version (middle row), and the 150K version (bottom row). As expected, precision decreases as size falls; this effect can be particularly observed in the most extreme experiment: the highest threshold (q =5) combined with the smallest dataset (150K version). On the other hand, the larger the dataset, the lower the recall results, since there will be more 'total relevant' items in each recommendation process. This effect is more severe when the threshold is not high (q =4), since even further 'total relevant' items will be in the denominator of the recall quality

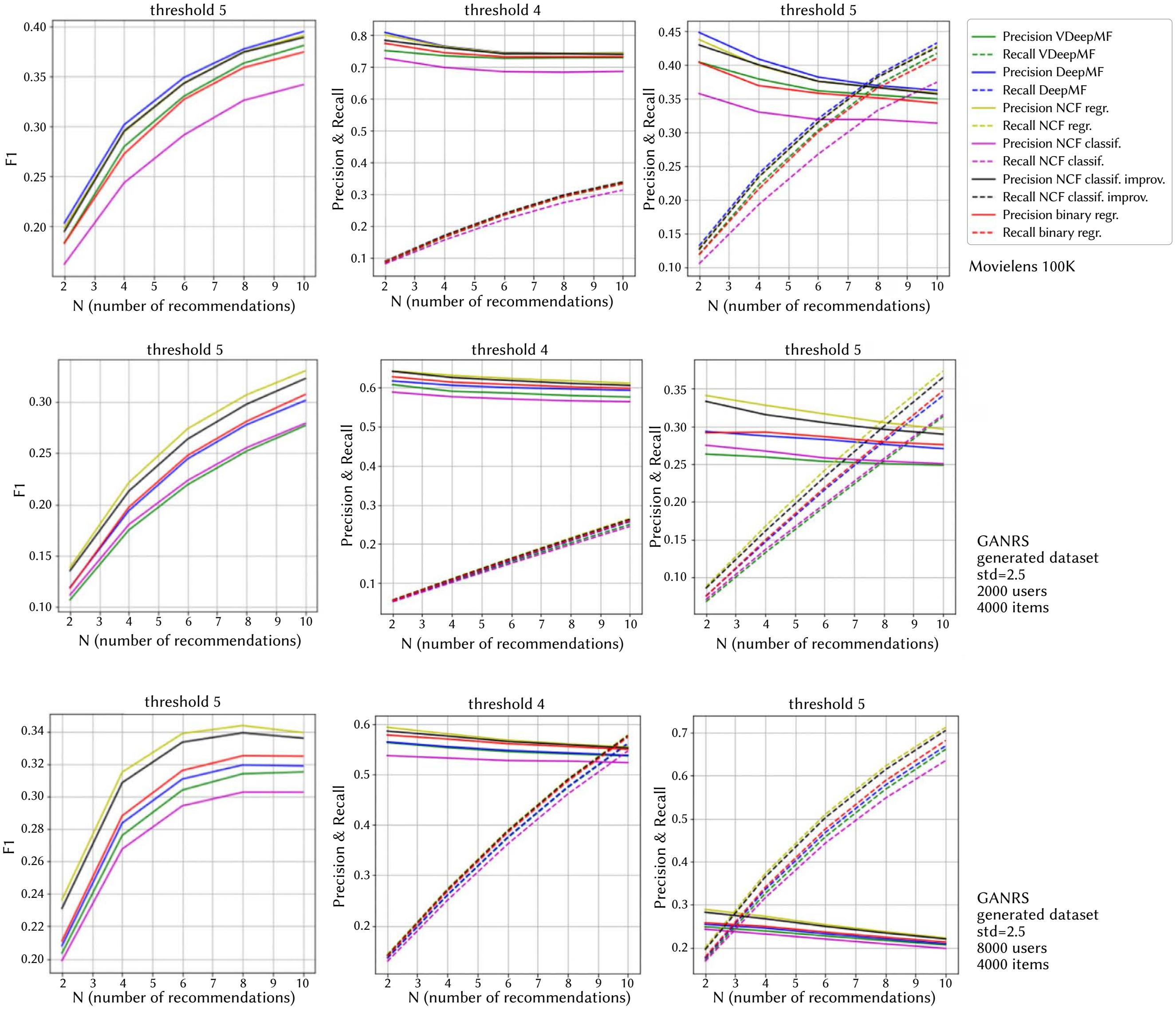


Fig. 6. Comparative among Movielens 100K, GANRS 2,000 users; and GANRS 8,000 users. Generated datasets with 4,000 items and 2.5 for the standard deviation of the Gaussian random noise. Number of recommendations N = [2, 4, 6, 8, 10].

measure. The top and middle graphs of the 'threshold 4' column in Fig. 5 show the concept. Beyond the specific quality values, we can observe that it is possible to use generated datasets with different sizes to test CF machine learning models in different scenarios: the results will show the expected behavior and trends. Regarding the tested deep learning models, it is interesting to observe how the NCF classification and, particularly, the binary regression dramatically decreases their performance when the dataset size increases. We can also see how the improved NCF classification reaches the NCF regression, compared to the results in Fig. 4.

C. Experiment 3: Movielens 100K Versus GANRS 2000 Users, Versus GANRS 8000 Users

This experiment compares the absolute values and the trends in the recommendation quality obtained for each baseline when applied to the source Movielens 100K dataset, to the GANRS generated dataset setting 2000 users, and to the GANRS generated dataset setting 8000 users. Both generated datasets take Movielens 100K as the source to catch its internal patterns. As in the previous subsections, we expect similar trends in the graph functions, showing that the GANRS generated datasets adequately mimic the Movielens 100K patterns. We also expect different absolute quality values due to the different number of users selected for each GANRS generated dataset.

To avoid unnecessary repetitions, experiments on the synthetic datasets generated from Movielens and MyAnimeList are restricted to the 2,000 versus 8,000 user comparatives.

Fig. 6 shows the Movielens results; they are similar to those obtained using generated datasets from Netflix*. In fact, both sets of synthetic data contain a similar number of samples: 405,539 versus 353,269 in the 2,000 user versions and 628,194 versus 509,193 in the 8,000 user datasets. Comparing the precision & recall results of the GANRS versions, both at thresholds 4 and 5 in Fig. 4 and Fig. 6, we can see that the absolute values (y-axis) and the curve trends are similar. Regarding the baselines, the NCF regression provides a balanced (F1) superiority, as it happens in the source Netflix* data set.

D. Experiment 4: MyAnimeList Versus GANRS 2000 Users, Versus GANRS 8000 Users

This experiment compares the absolute values and the trends in the recommendation quality obtained for each baseline when applied to the MyAnimeList dataset, to the GANRS generated dataset setting 2000 users, and to the GANRS generated dataset setting 8000 users.

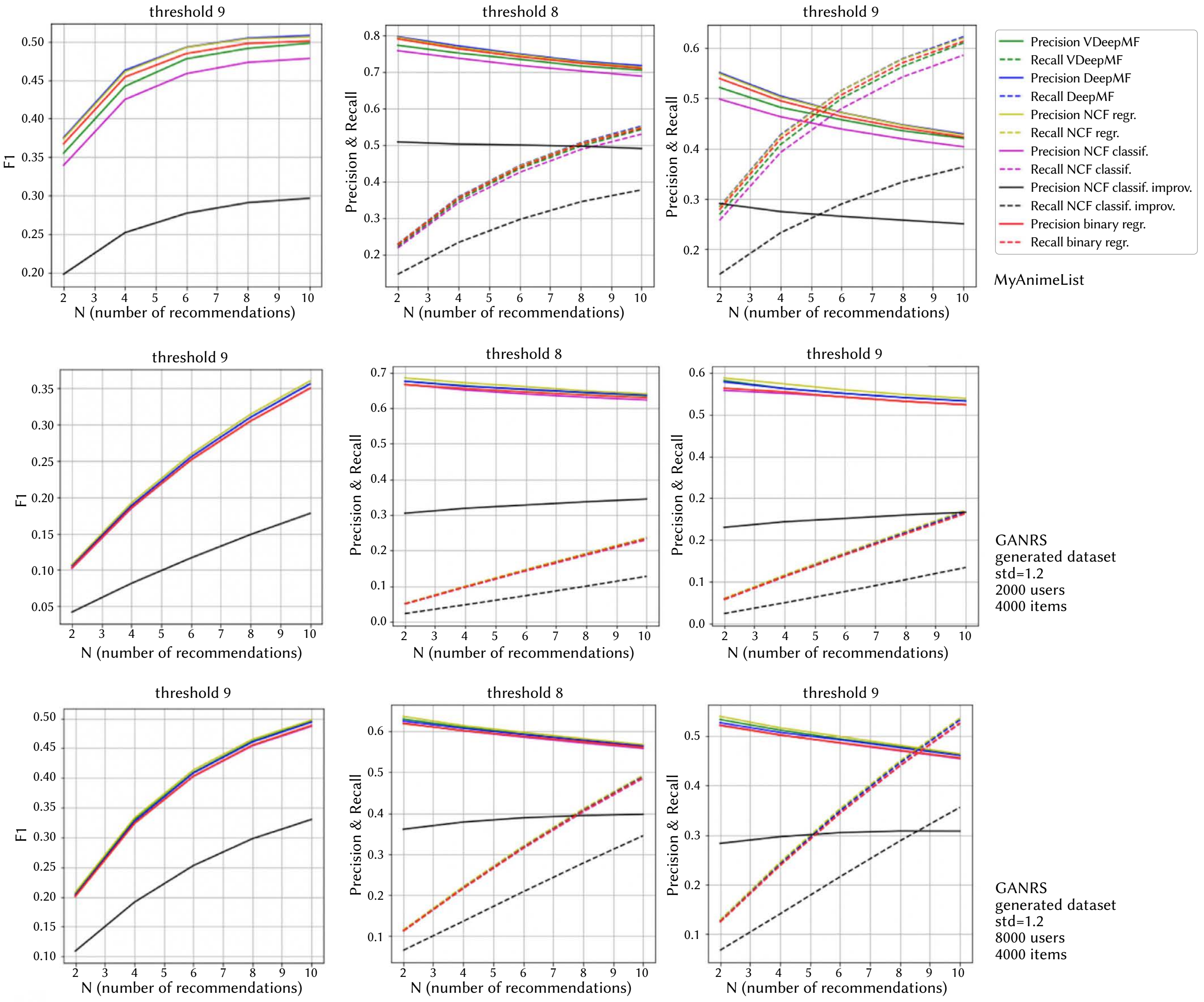


Fig. 7. Comparative among MyAnimeList, GANRS 2,000 users; and GANRS 8,000 users. Generated datasets with 4,000 items and 2.5 for the standard deviation of the Gaussian random noise. Number of recommendations N = [2, 4, 6, 8, 10].

The MyAnimeList family of generated datasets provides interesting results, since MyAnimeList contains a range of ten ratings (1 to 10) instead of the usual 1 to 5. Focusing on the threshold (*q*=10) in Fig. 7, it can be observed that precision improves (compared to the preceding results when *q*=5). It probably happens due to a higher proportion of ratings 10, compared to the equivalent (ratings 5) in Movielens or Netflix*. The important here is that the synthetic datasets in Fig. 7 mimic this behavior; that is: the comparative between MyAnimeList (Fig. 7 top right graph) and Movielens/Netflix* (Fig. 4 and Fig. 6 top-right graphs), looks similar to the comparative between the MyAnimeList GANRSs (Fig. 7 middle-right and bottom-right graphs) and Movielens/Netflix* GANRSs (Fig. 4 and Fig. 6 middle-right and bottom-right graphs). This means that the GANRS synthetic datasets are adequate. Finally, as expected, the classification models perform worst in this scenario (exception the improved one), since it is harder to correctly classify ten categories than five categories.

IV. Discussion

Overall, the obtained results show that the synthetic GANRS datasets adequately mimic the behavior of the source datasets from which the GAN learns their patterns. Results sustain the hypothesis of the paper, and they confirm that the GANRS generator creates synthetic datasets containing similar patterns and probability distributions to the chosen source datasets, and what is more: this is also true when the selected number of users, items and samples varies. Our view is that the GANRS generative model gets its successful behavior from the architectural key with which it has been designed: to feed the GAN kernel of the model with short and dense embeddings instead of the traditional large and sparse raw data [15]. In this way the GAN stage improves its performance, better catches the source patterns, and it reduces the mode collapse condition.

Since the synthetic datasets can be generated setting their sizes, number of items, and number of users, it is possible to use them to test CF machine learning models on different scenarios, e.g., when the number of users varies. Specifically, all the synthetic datasets tested in the experiments show adequate variation of precision and recall, where precision improves (and recall gets worst) as the number of samples increases. This is because the higher the total number of samples, the higher the average number of ratings for each user. Additionally, as expected, accuracy and recall differ when tested by setting different recommendation thresholds. Observing the results of the tested deep learning models, the NCF regression and the improved NCF classification perform significantly better than the NCF classification and the binary regression. DeepMF and VDeepMF provide slightly lower quality results than NCF regression. All these results are compatible with the state-of-the-art ones. Finally, it is remarkable how the tested GANRS datasets adequately catch the quality loss of the NCF classification when MyAnimeList is taken as a source, since this dataset encodes a ten ratings interval instead of the usual five ratings interval.

V. Conclusions

This paper tests the performance of the synthetic datasets generated from the published GANRS method. A representative set of generated datasets has been created by selecting a different number of users and a different number of samples. The obtained datasets have been tested on six representative CF deep learning models: DeepMF, VDeepMF, NCF, NCF classification, improved NCF classification, and binary regression. The recommendation quality measures precision, recall, and F1 have been chosen. The results show adequate performance of the synthetic datasets on all applied deep learning models. In particular, it can be observed that, as expected, precision improves when the size of the dataset increases, as well as when the average number of ratings of each user also increases. In the same way, the recall decreases as the size of the data set increases. The interval of the ratings in the dataset (ten in MyAnimeList and five in Movielens and Netflix*) has the expected impact, where both the recall and, particularly, the precision drop using MyAnimelist. The tested CF deep learning models perform similarly when the results of the synthetic datasets are compared with the real datasets, and it happens on the different combinations of the selected number of samples and number of users. Overall, the GANRS method generates valuable synthetic datasets that can be used to test new deep learning models proposed in the CF RS area. Future works include testing synthetic datasets tailored to specific CF scenarios such as user cold start, item cold start, dataset cold start, imbalanced data, demographic variations, binary ratings (like, non-like), fairness, recommendation to groups of users, and heavy sparse data.

Acknowledgment

This work was partially supported by Ministerio de Ciencia e Innovación of Spain under the project PID2019-106493RB-I00 (DL-CEMG) and the Comunidad de Madrid under Convenio Plurianual with the Universidad Politécnica de Madrid in the actuation line of Programa de Excelencia para el Profesorado Universitario.

References

[1] J. Bobadilla, A. González-Prieto, F. Ortega, R. Lara-Cabrera, "Deep learning feature selection to unhide demographic recommender systems factors," *Neural Computing and Applications*, vol. 33, no. 12, pp. 7291-7308, 2021.

[2] Y. Deldjoo, M. Schedl, P. Cremonesi, G. Pasi, "Recommender systems leveraging multimedia content," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1-38, 2020.

[3] S. Kulkarni, S.F. Rodd, "Context aware recommendation systems: A review of the state of the art techniques," *Computer Science Review*, vol. 37, 100255, 2020.

[4] J. Shokeen, C. Rana, "A study on features of social recommender systems", *Artificial Intelligence Review*, vol. 53, no. 2, pp. 965-988, 2020.

[5] J.B. Schafer, D. Frankowski, J. Herlocker, S. Sen, "Collaborative Filtering Recommender Systems," in: The Adaptive Web. *Lecture Notes in Computer Science*, Brusilovsky, P., Kobsa, A., Nejdl, W. (eds), Springer, Berlin, Heidelberg, 2007, vol. 4321.

[6] E. Cano, M. Morisio, "Hybrid recommender systems: A systematic literature review," *Intelligent Data Analysis*, vol. 21, no. 6, pp. 1487-1524, 2017.

[7] B. Zhu, R. Hurtado, J. Bobadilla, F. Ortega, "An efficient recommender system method based on the numerical relevances and the non-numerical structures of the ratings," *IEEE Access*, vol. 6, pp. 49935-49954, 2018.

[8] A. Mnih, R. R. Salakhutdinov, Probabilistic matrix factorization, *Advances in neural information processing systems*, vol. 20, 2007.

[9] C. Févotte, J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural computation*, 2011, vol. 23, no. 9, pp. 2421-2456, 2011.

[10] H.-J. Xue, X. Dai, J. Zhang, S. Huang, J. Chen, "Deep Matrix Factorization Models for Recommender Systems," in: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017, pp. 3203-3209.

[11] J. Bobadilla, J. Dueñas, A. Gutiérrez, F. Ortega, "Deep Variational Embedding Representation on Neural Collaborative Filtering Recommender Systems," *Applied Sciences*, vol. 12, no. 9, 4168, 2022.

[12] X. He, L. Liao, H. Zhang, "Neural Collaborative Filtering," *International World Wide Web Conference Committee (IW3C2)*, 2017, pp. 173-182.

[13] J. Bobadilla, F. Ortega, A. Gutiérrez, S. Alonso, "Classification-based Deep Neural Network Architecture for Collaborative Filtering Recommender Systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 1, pp. 68-77, 2020.

[14] J. Bobadilla, A. Gutiérrez, S. Alonso, A. González-Prieto, "Neural Collaborative Filtering Classification Model to Obtain Prediction Reliabilities," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, pp. 18-26, 2022.

[15] J. Bobadilla, A. Gutiérrez, R. Yera, L. Martínez "Creating Synthetic Datasets for Collaborative Filtering Recommender Systems using Generative Adversarial Networks," *Knowledge-Based Systems*, pre-proof: 111016, 2023. https://doi.org/10.1016/j.knosys.2023.111016.

[16] M. Gao, J. Zhang, J. Yu, J. Li, J. Wen, Q. Xiong, "Recommender systems based on generative adversarial networks: A problem-driven perspective," *Information Sciences*, vol. 546, pp. 1166-118, 2021.

[17] Y. Deldjoo; T. Noi, F.A. Merra, "A Survey on Adversarial Recommender Systems: From Attack/Defense Strategies to Generative Adversarial Networks," *ACM computing surveys*, vol. 54, no. 2, pp. 1-38, 2021.

[18] D.-K. Chae, J.-S. Kang, S.-W. Kim, J.-T. Lee, "CFGAN: a generic collaborative filtering framework based on generative adversarial networks," in: Proceedings of the 27th, *ACM International Conference on Information and Knowledge Management*, CIKM 2018, 2018, pp. 137-146.

[19] G. Guo, H. Zhou, B. Chen, et al., "IPGAN: Generating informative item pairs by adversarial sampling," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no.2, pp. 694-706, 2022.

[20] H. Bharadhwaj, H. Park, B.Y. Lim "Recgan: recurrent generative adversarial networks for recommendation systems," in: *Proceedings of the 12th ACM, Conference on Recommender Systems*, RecSys 2018, 2018, pp. 372-376.

[21] J. Zhao, H. Li, L. Qu, Q. Zhang, Q. Sun, H. Huo, M. Gong, "DCFGAN: An adversarial deep reinforcement learning framework with improved negative sampling for session-based recommender systems," *Information sciences*, vol. 596, pp. 222-235, 2022.

[22] J. Wen, X. Zhu, C.D. Wang, Z. Tian, "A framework for personalized recommendation with conditional generative adversarial networks," *Knowledge and information systems*, vol. 64, no. 10, pp. 2637-2660, 2022.

[23] J. Sun, B. Liu, H. Ren, W. Huang, "NCGAN: A neural adversarial collaborative filtering for recommender system," in: *Journal of intelligent & fuzzy systems*, vol. 42, no. 4, pp. 2915-2923, 2022.

[24] W. Shafqat, Y.C. Byun, "A Hybrid GAN-Based Approach to Solve Imbalanced Data Problem in Recommendation Systems," in: *IEEE access*, vol. 10, pp. 11036-11047, 2022.

[25] M. Mladenov, C.W. Hsu, V. Jain, E. Ie, C. Colby, N. Mayoraz, H. Pham, D. Tran, I. Vendrov, C. Boutilier, "Demonstrating Principled Uncertainty Modeling for Recommender Ecosystems with RecSim NG," in: RecSys 2020 - *14th ACM Conference on Recommender Systems*, 2020, pp. 591–593.

[26] J.C Shi, Y. Yu, Q. Da, S.Y. Chen, A.X. Zeng, "Virtual-Taobao: Virtualizing real-world online retail environment for reinforcement learning," in*:* *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 33, no. 01, 2019, pp. 4902–4909.

[27] M. del Carmen, S. Ilarri, R. Hermos, R. Trillo-Lado, "Datagencars: A generator of synthetic data for the evaluation of contextaware recommendation systems," *Pervasive and Mobile Computing*, vol. 38, pp.

516–541, 2017.

[28] V. Provalov, E. Stavinova and P. Chunaev, “SynEvaRec: A Framework for Evaluating Recommender Systems on Synthetic Data Classes,” in: *International Conference on Data Mining Workshops (ICDMW)*, Auckland, New Zealand, 2021, pp. 55-64.

[29] F. Ortega, B. Zhu, J. Bobadilla, A. Hernando, “CF4J: Collaborative filtering for Java,” *Knowledge-Based Systems*, vol. 152, pp. 94-99, 2018.

[30] M. Račinský, “MyAnimeList Dataset,” Kaggle, 2018. [Dataset]. Available: https://www.kaggle.com/azathoth42/myanimelist, doi: 10.34740/KAGGLE/DSV/45582.

[31] F.M. Harper, J.A. Konstan, “The movielens datasets: History and context,” *ACM Transactions on Interactive Intelligent Systems*, vol. 5, no. 4, pp. 1-19, 2015.

Jesús Bobadilla

Jesús Bobadilla received the B.S. and the Ph.D. degrees in computer science from the Universidad Politécnica de Madrid and the Universidad Carlos III. Currently, he is a full professor with the Department of Information Systems, Universidad Politécnica de Madrid. He is a habitual author of programming languages books working with McGraw-Hill, Ra-Ma and Alfa Omega publishers. His research interests include information retrieval, recommender systems and speech processing. He oversees the FilmAffinity.com research team working on the collaborative filtering kernel of the web site. He has been a researcher into the International Computer Science Institute at Berkeley University and into the Sheffield University. Head of the research group.

Abraham Gutiérrez

Abraham Gutiérrez received the B.S. and the Ph.D. degrees in computer science from the Universidad Politécnica de Madrid. Currently, he is currently an associate professor with the Department of Information Systems, Universidad Politécnica de Madrid. He is the author of search papers in most prestigious international journals. He is a habitual author of programming languages books working with McGraw-Hill, Ra-Ma and Alfa Omega publishers. His research interests include P-Systems, machine learning, data analysis and artificial intelligence. He is in charge of this group innovation issues, including the commercial projects.

Explainable Artificial Intelligence-Based Diseases Diagnosis From Unstructured Clinical Data and Decision Making Using Blockchain Technologies

Sumathi M.[1]*, S.P. Raja[2]

[1] School of Computing, SASTRA Deemed University, Thanjavur, Tamil Nadu (India)
[2] School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, (India)

* Corresponding author: sumathi@it.sastra.edu

Received 30 September 2022 | Accepted 28 December 2022 | Published 6 February 2025

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

In the digital era, health information is stored in digital form for easy maintenance, analysis and transfer. The proficiency of manual illness diagnosis and drug prediction in the medical field depends on the expertise availability, and experience of the specialists. In emergency and abnormal situation, the patient's life completely depends on expert's availability. Therefore, a different approach is needed to get around the difficulties in managing emergency cases. Artificial intelligence helps to take decisions in an accurate manner but does not provide the details of the decisions. The ability to treat emergency patients entirely depends on the particular hospitals. The clinical data includes numerical results, text prescriptions, scanned images, etc. Therefore, managing unstructured data with care is necessary for making clinical decisions. An explainable artificial intelligence-based disease diagnosis and blockchain-based decision-making system are presented in this work to address these challenges and improve patient care. A natural language processing system analyzes the unstructured data to identify different types of data and explainable AI diagnosis disease with justification and reason for the prediction. An ant colony optimization-based recommender system examines the predicted decision and identifies the specific drug for the disease. The disease decision and drug information are kept in a permissioned blockchain for confirmation. Decisions are validated by more than 50% of the experts present in the permissioned blockchain network, which consists of experts from various regions. As a result, the quickest and most accurate decisions possible are taken to handle emergency situations.

Keywords

Ant Colony Optimization, Blockchain, Clinical Care, CNNn-LSTM, Disease Diagnosis, Explainable AI, Unstructured Data.

DOI: 10.9781/ijimai.2025.02.003

I. Introduction

In the past, medical experts had examined minimal sized clinical data (CD) sets for disease predictions (DP) and drug recommendations. At present, automation devices produce an enormous amount of CD as a result of technological development. To analyse these huge amounts of data by an expert is not an easy task and requires more time to predict the diseases and drugs. It leads to error-prone in emergency handling cases. Currently, artificial intelligence (AI) predictions are favoured in CD analysis [1]. The capability of the training tool and model determine how accurate the current AI predictions are. But, the diagnosis (Dig) is unclear based on the predicted results of current AI, and it is challenging to manage the enormous volume of crucial data. As a result, the present AI has significant rates of false positives. Consequently, a new tool with enhanced functionality is needed. The vast majority of the CD is composed of tiny critical terms, and these tiny critical terms are crucial for disease Dig (DDig). Consequently, the work of significant phrase extraction is crucial in DDig [2].

Natural language processing (NLP) is a well-known method for gleaning tiny critical terms from a vast amount of structured, semi-structured, and unstructured CD. Hence, NLP is a useful tool for handling diverse data effectively. The CD includes text, numbers, scanned images, handwritten reports, and audio data. NLP extracts the tiny critical terms which have been required for decision-making (DM) from the diverse structured CD [3]. AI helps to analyse the large volume of digitalized medical information in a quicker way and produces the result without any explanation. This prediction is not an evident result for making a decision. Hence, the explainable AI (XAI) is preferable for making accurate results with evident. In an XAI, the results are discussed with justification and provide detailed results of the analysis. By using this result, the important features are analysed and better decisions are taken for a disease. Hence,

Please cite this article as:
Sumathi M., S.P. Raja. Explainable Artificial Intelligence-Based Diseases Diagnosis From Unstructured Clinical Data and Decision Making Using Blockchain Technologies, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 40-51, 2025, http://dx.doi.org/10.9781/ijimai.2025.02.003

XAI has been used in the proposed system. In healthcare domain, recommender system (RCS) has been used for improving clinical care and provides suggestions in different aspects like disease prediction, drug recommendation and medical information storage etc. This RCS helps patients and clinical care takers for taking accurate decisions in normal and emergency situations. To improve clinical care, the ant colony optimization (ACO) based RCS has been used in the proposed system. The ACO technique compares the historical disease and drug predication result with the new prediction results and produces the result as "recommended" or "not recommended".

Presently, blockchain (BC) technology is used in different domains for the sharing of information in an immutable and decentralized way. In a medical sector, sharing of information between experts in different geographical location is an essential task of handling the critical and emergency cases. The clinical information requires highly secure and immutable sharing. Thus, the BC technique is used in the proposed system. In a proposed system, the BC technology helps to exchange confidential data between AI and clinical professionals in a secured way without third parties. The property of BC is to provide decentralized and irreversible storage. Hence, the BC technique is used for transferring the predicted diseases and recommended drug information to a clinical expert in a confidential and immutable way [4]. The information stored in the block is validated by the clinical experts. If it is an accurate prediction, experts "approve to the prediction" otherwise "reject the prediction". Information about approved diseases and drug information is stored in BC. The information stored in BC cannot be changed in future. This immutable storage is useful for future DM [5].

This article's remaining portion is organized as follows: In section 2, the methods now in use and the advantages and disadvantages of AI techniques have been discussed. In section 3, along with the essential equations and architecture, the proposed NLP data classification, XAI disease Dig, ACO drug prediction and BC based DM framework principles and algorithms are described. The experimental setup, dataset gathering, and suggested methodology for experimental outcomes in various parametric aspects are all covered in section 4. The proposed system security strength is covered in section 5 along with a proof of concept. In section 6, the proposed system finishes with future improvements.

II. Literature Review

A. Term Extraction Using NLP

Jignesh R.Parikh et al. have employed NLP to find rare diseases. Correlating genetic variation data with in-depth phenotypic data is necessary for effective genetic Dig. The manually extracted terms and the extracted NLP terms have been associated with preparing a rank list of the terms. Then, the higher ranked terms have been filtered out for analysis [6]. Julia Lve et al. had concentrated on the most challenging of de-identifying and analyzing the free text in CD. The generated text from a subsequent NLP text categorization was used for conducting the extrinsic evaluation. The important terms of the CD were selected, and fictional data was created for accessing the test data [7]. Essam H. Houssein et al. had combined NLP and AI to extract the information from the CD. The biomedical NLP helps for direct DM and has been used for unlocking the hidden CD content. Additionally, analyzing CD and extracting previously undiscovered CD is possible in biomedical NLP [8]. Comito et al. have been discussed laboratory test results, patients' basic information, CD, and social media data for their heterogeneous health data integration. In order to forecast the patient's future health information, a neural network model was deployed. The similar technique has been used to predict the similar diseases. Based on similarities, the combined supervised and unsupervised techniques with NLP were utilized to forecast the diseases [9].

Sungho Sim et al, had discussed how existing IoT and AI techniques have been worked together to Dig and monitor diseases in an emergency case. This work specifically discussed the rapid spread of infectious diseases. To determine the severity of the disorders, thorough analysis is necessary for disease prediction [10]. Haohui Lu et al. have discussed regular patient monitoring is an expensive and impractical chore. As a result, the bipartite graph has been used for framing the patient network among the patients who were all Dig with the same condition. To estimate the disease risk, network analysis and machine learning (ML) algorithms were integrated. Random forest (RF) offers higher prediction accuracy than other ML methods [11]. Flora Amato et al. collect the CD from the patients using IoT smart devices, however evaluating this data is a challenging task because data is collected from numerous devices. This problem was solved by extracting structured data (SD) from the CD and processing it semantically. The data was stored on a central server to facilitate various categorization and DM techniques. NLP were used to extract the information from CD's and data from smart devices. The feature extraction (FE) and reduction reduces the data size and enables us to handle large-sized data [12].

B. Disease Diagnosis Using AI

The objective of the Methods section is to describe materials and methods in a detailed way so that a knowledgeable reader could repeat the experiment. Possible sub-sections could be: *Participants, Materials, Tasks* and *Design and Analysis.* Notice that not all these sections are always applicable.

Eden et al. had been analyzed, the diabetes is the root cause of a number of medical problems, including heart attack, stroke, renal and heart failure, and coma. Diabetes is brought on by genetics, environmental circumstances, immune system attacks and insulin breakdown, immune system responses to an infection, among other things. The AI prediction is useful for forecasting the disease severity accurately. Applications of AI in drug therapy include, improving drug treatment procedures and foreseeing drug-drug interactions [13]. Elham Khodabandehloo et al. were discussed the early DDig to lessen the impact of disease severity. The AI-enabled smart house alerts practitioners for providing better care for elderly. In the present system, the forecast had been made without any justification. The practitioner recommended the rule-based prediction system for delivering better results. The rule-based approach is unable to handle a variety of users and abnormal events. Thus, the author has suggested flexible AI to identify the initial symptoms and provide a fine-grained justification for forecasts. This adaptable, data-driven XAI can be adjusted to various people and circumstances. A collaborative process has been used for anomaly identification. The author's concentrated on fixed parameters, quick analyses, and DM on specific criteria. The multi-criteria and long-term study has not been covered in this work [14].

Thomas et al. discussed about the four main factors that affect the accuracy rate of AI models. The data source had been considered as the primary factor. The CD gathered from various sources may be out-of-date, biased, incorrect, and lacking. As a result, false Dig had been made. However, both people and information sources are important to take accurate decision. The bias of AI Dig is the second factor. The training data or human classification may have skewed the AI Dig. Hence, the properties of the dataset and categorized by humans should be examined before moving forward with the Dig model's finalization. AI Dig analysis is considered as the third factor. The performance of the locked AI model depends on the first learning set's initial value. It results in errors, over- Dig or under Dig for certain patients. Therefore, before making a forecast, it is necessary to assess the performance of an AI. The way that Dig work is organized and divided can also have an impact on the accuracy of the Dig. Therefore, prior to Dig the legal obligation for labor must be established [15].

Norma Latif Fitriyani et al. had been discussed the CD with normal, abnormal and outlier data. The outlier data must be taken out of the dataset before it can be used to make predictions. The outlier data had been found and eliminated by DBSCAN method. The unbalanced training dataset was balanced by SMOTE-ENN, and the prediction model was learned and created using XGBoost ML algorithm. To verify the effectiveness of the AI model, the gathered patient data has been compared with statistical significance [16]. Romany Fouad Mansour et al. had been discussed an automated prediction that integrates the IoT and AI in order to make a decision quickly. The Crow Search Optimization algorithm-Cascaded Long Short Term Memory (CSO-CLSTM) model was used for DDig. The weight and bias settings of the CLSTM technique were adjusted using the CSO approach. The outliers have been eliminated by an isolated forest technique. The accuracy of the CSO-LSTM approach was Dig diabetes and heart disease 96.16% and 97.26%, respectively [17]. Vijendra Singh et al. had been discussed the ML based hybrid classification technique for DDig. The support vector machine (SVM) and dimensionality reduction methods were combined to identify the chronic kidney disease. Compared to existing techniques, this combination provided higher accuracy rate [18].

C. Recommendation System-Based Drug Prediction

Dinakaran et al. had discussed the spectral deep feature classification-based drug RCS. The drug ratio is determined by feature selection and mutual molecular weights. The characteristics are using the candidate selection technique, and the molecular weight has been determined by the intra-class activation function. Then, the suggested drug level had been determined using the active adaptive recurrent neural network. More than 94% of the predicted accuracy has been produced by the RCS [19]. Farman Ali et al. were proposed an ontology-based RCS for IoT-based healthcare. The Protégé web ontology language had been used to predict the patients' dietary habits, medical history, and medication needs. Wearable health monitoring devices were used to monitor the data on the patient's body. Then, the RCS had been automated by fuzzy logic and semantic web rule languages. Without any preparation, this procedure performs straight analysis of the data. As a result, the RCS's accuracy rate suffers [20]. Farman Ali et al. were proposed the wearable sensors and social networking data based health monitoring system. Bidirectional long-short-term memory had been utilized to predict abnormal conditions, including high blood pressure, diabetes, mental health issues, and medication adverse effects. It offered a greater accuracy rate in the health monitoring system and handled diverse data [21].

Luis Fernando Granda Morales et al. had discussed the clustering and collaborative filtering based drug RCS for diabetes. The dimensionality reduction and clustering had been performed to predict the drug for diabetes. The collect user profile was compared to other patient profile for finding the similar characteristics. At last, the RCS group the patients based on the identified values. Then the quality of the recommendation and predictions were compared to the benchmark data for evaluating the accuracy rate [22]. Qing et al. had been created the unified framework with RCS and knowledge graph. The basic low-dimensional entities were learned from knowledge graph and integrated this result in neural factorization machine for improving accuracy result. Compared to existing RCS, the integrated RCS provided higher prediction accuracy [23].

D. Blockchain-Based Clinical Data Storage

Nowadays BC technology used in different sectors like agriculture, medical, manufacturing etc. Guofeng et al. had used the BC technique in agriculture to store the information in a secured and fine-grained accessible manner. Both horizontal and vertical partitions are applied to partition the data and stored in a block instead of actual data. This partition based technique provides higher confidentiality of sensitive information. Additionally, the cipher policy attribute based encryption and data encryption standard techniques were used for providing data confidentiality to block information [24]. Tetiana Hovorushchenko et al. had discussed the method for blockchain based medical data storage. The header node contains the date, time, version, metadata, digital signature, encrypted data and hash code of the previous block etc. By the same way, the next blocks were created and added to the existing blocks. This technique provides integrity to user data [25].

Suruchi et al. had reviewed the BC technology in healthcare applications. By using a BC network, patients CD from various organizations can be stored in a single network, which makes it easier for patients, doctors and other caretakers to make accurate decisions. The BC method offers patients CD protection and immutability [26]. Roberto Cerchione et al. were proposed the BC-based CD storage. To provide higher security, the CD had been kept in the permissioned BC. By using an information processing principle, the CD can be verified and transferred to storage. This procedure reduces processing and storage expenses, medical errors, and raises service standards. As a result, the BC approach contributes to enhancing patient care in a safe manner [27]. Gaofan Lin et al. had stored CD and offered fine-grained data sharing amongst caregivers using the BC approach. To guarantee privacy for sensitive information, the encrypted data had been kept on the block. Caretakers gave secure fine-grained access through the deployment of the ciphertext searching. As a result, patient data had been kept confidential and accurate [28].

E. Summary

- NLP has been used for extracting critical CD from the diverse data in an efficient manner. The existing works concentrate on specific data type not to diverse data.
- XAI is used for DDig with a justification of the prediction. This feature helps to take accurate decision making than the AI techniques. Only limited features are analyzed in the existing works.
- The RCS-based drug prediction helps to predict the drug in a fast and accurate manner.
- The BC technique helps to store the medical data in a secure and immutable way. The inter-organization based network formation is not yet designed.

Based on the analysis provided above, the unstructured CD requires special care to DDig and drugs prediction. The main contribution of the proposed effort is outlined below in order to meet these needs:

III. Contribution of Proposed Work

- NLP is used for assessing the enormous amount of unstructured data (USD) gathered from many sources with varying levels of scarcity. The preferred phrases have been extracted from the structured, semi-structured, and unstructured CD by utilizing medical code generation (MCG), pattern matching (PM), keyword extraction (KE), and visual feature extraction (VFE) approaches.
- Similarity mapping is used for designing the highly efficient XAI training model (for providing explanation and understanding of DDig). By using single mapping process, the 100% similarity mapping is an impossible task. Hence, the multi-level training model is designed for reaching the nearly 100% prediction accuracy.
- The ACO RCS performed the depth-first search on the predicted diseases based on historical data for predicting the best drug for the disease. The ACO RCS provides higher clinical care to the patients on time.

- By using BC network, the reliable and secured communication channel has been created between AI and clinical specialists. The blocks in BC stores CD securely, and helps to validate prediction results while managing emergency situations. Based on accepted prediction, the clinical professionals construct and add the safe and immutable blocks in the BC network. This immutable storage will be beneficial for future prediction.

IV. Proposed Work

Currently, health data is gathered from a variety of sources, including IoT sensor devices, CD, narrative descriptions from clinical experts, patient basic information, and scanned images. It is a challenging task to manage this enormous data in its original form and making decisions on DDig. Consequently, the DDig and drug prediction leads to inaccurate DM. In order to address these problems, the enormous USD must be transformed into SD. Among the numerous data, only minimal size data is useful for making predictions about the nature of diseases, their severity, etc. To increase prediction accuracy and to provide better clinical care, it is necessary to identify and extract the necessary terms of the obtained data. Recently AI has been utilized in the healthcare industry to accurately predict diseases. Diverse AI technologies are being developed by researchers and AI specialists to increase the precision of DDig, promote drug development, provide more advanced patient monitoring, and cater to the needs of the person.

The AI tools help to reduce mistakes in manual DDig. Currently, AI models are predicting diseases without clear justification (no clear justification for why the condition is predicted). Consequently, it is challenging to grasp DDig. Likewise, the current AI techniques analyze the diseases using similar values, and it is hard to map with 100% similarity. As a result, these technologies aim to increase the accuracy of DDig and similarity mapping. The dissimilar values are not analyzed or not taken into account for further processing. Poor prediction in DDig is leading to unclear and different results. To increase the accuracy of the prediction, to design a new AI tool for identifying the diseases with a clear understanding is required.

In order to offer better clinical care, a clinical expert must get the drug RCS based on the DDig. The RCS suggests a drug based on an analysis of previous drug information for the same disease. Depending on the type of RCS, the RCS accuracy rate varies. The previous drug suggestion success rate must be examined and projected in order to establish the current drug RCS and increase RCS accuracy. Currently, a third-party services (CSP) using cloud storage site to store and maintain the predicted CD. This CSP leads to security breaches, including erroneous RCS and adversary (Ã) access or avoiding authorized user access. A novel storage approach with security, integrity, and availability is required to ensure data privacy and security of CD. In Fig. 1, the suggested work has been displayed.

The proposed method uses cosine similarity algorithms, PM, and sentence matching (SM), word embedding (WE), KE, and MCG to extract the necessary terms from the collected data CD. Following that, XAI-based multi-level Convolution Neural Networks (CNNn) and local interpretable model-agnostic explanations (LIME) algorithms are used to match the retrieved phrases. To map all the extracted terms to the trained model at multi-level, an integrated model has been created. By this mapping, nearly 100% similarity and accuracy of predictions is achieved. The ACO has been used to identify drugs after DDig in order to improve clinical care for patients. Finally, the BC has been used for reliable and secure data transmission between clinical experts and AI for transferring DDig and drug RCS. CD is securely and immutably stored in BC, making it helpful for upcoming DM.

The clinical professionals that work in various healthcare institutions have involved the BC network. The DDig and drug RCS are verified by the clinical experts in the BC network. If the prediction is accurate, the block has been created and added to the BC network. This procedure increased the suggested technique's prediction accuracy. Likewise, this network effectively handles emergency situations. The accuracy of clinical care and DDig has been enhanced by the suggested technique.

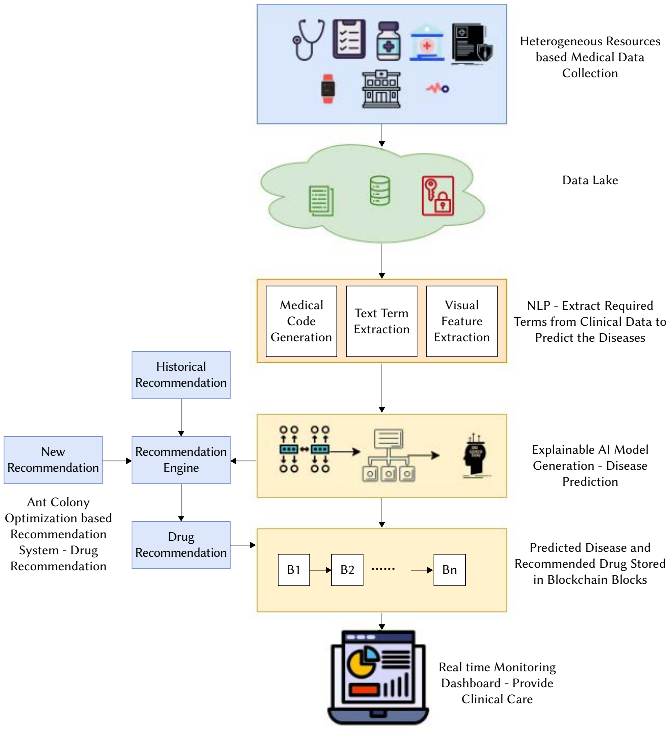

Heterogeneous Resources based Medical Data Collection
Data Lake
Medical Code Generation
Text Term Extraction
Visual Feature Extraction
NLP - Extract Required Terms from Clinical Data to Predict the Diseases
Historical Recommendation
New Recommendation
Recommendation Engine
Explainable AI Model Generation - Disease Prediction
Ant Colony Optimization based Recommendation System - Drug Recommendation
Drug Recommendation
B1
B2
......
Bn
Predicted Disease and Recommended Drug Stored in Blockchain Blocks
Real time Monitoring Dashboard - Provide Clinical Care

Fig. 1. Flow Diagram of the Proposed Work.

V. Methodology

A. Term Extraction From Clinical Record Using NLP

From free hand texts, images, SD, IoT sensor data, insurance providers, social media, and web knowledge, the CD are extracted. Diverse data scarcities are present in the data gathered from diverse sources. To extract relevant terms from this variety leads to high processing complexity. Therefore, the USD is converted into SD by using systematized nomenclature of medicine (SNOMED), international classification of diseases (ICD-CM) codes. Since, the SD analysis and validations are simpler than USD process. NLP helps to fix this issue. The simplest way to turn USD CD into SD is by using the NLP. The clinical caretakers benefit from the NLP conversion process in a variety of ways, including the reduction of manual analysis time, the production of precise and effective solutions, the provision of safety reviews, the ability to handle large-scale automated processing, and the management of large volumes. To transform USD into SD, various methods have been applied. Here is a list of them:

1. Medical Code Generation

The USD has been transformed into medical codes (MC) like unified medical language system (UMLs), ICD, and SNOMED for faster processing, risk and disease veracity prediction. MC is the process of converting diagnoses, treatments, services, and equipment used in healthcare into standard medical alphanumeric codes. Medical record material, like transcriptions, laboratory and radiologic results has been used to generate the Dig and procedure codes. The Health Care Procedural Coding System (HCPCS) level II categorization system and ICD-10-CM has been used to examine the USD CD and assign standard codes. The CD lists more than a thousand health conditions, diseases, injuries, and other causes. The classification, reporting, and tracking are made simpler by the usage of medical coding techniques (MCT). Every disease has various descriptors, acronyms and eponyms used in the healthcare industry. The MCT effectively standardizes the language and presentation of each component, making it simpler to comprehend, analyze, track and modify. Punctuation, special characters, mathematical symbols, and URLs are all have been deleted from USD during preprocessing. The stop-words have been eliminated for tokenization. The ICD-10 code book has been used to convert the key terms into MC. XAI uses this coded data to forecast diseases. Table I. Shows the sample ICD-10 code and its description.

TABLE I. Sample ICD-10 Code Format [29]

Code	Description
E08.220	Diabetes
E09.520	Drug or chemical
E10.110	Type 1 diabetics
E11.410	Type 2 diabetics
125.110	Arteriosclerotic heart disease
K50.013	Crohn's disease
L89.213	Pressure ulcer of right hip, stage III

2. Text Extraction From Semi-Structured Text Document

The PDF CD has been taken as the input into a semi-structured document (SSD). Equation (1) has been used for term extraction (TE). The semantic meaning of the text sequence has been captured by the word embedding technique.

$$\mathrm{Term\ Extraction}(\mathrm{T_E}) = \begin{cases} \mathrm{P_M} \leftarrow \mathrm{P_{ID}}, \mathrm{Ph_{no}}, \mathrm{Age} \\ \quad \mathrm{S_M} \leftarrow \mathrm{P_{name}}, \mathrm{P_{Location}}, \mathrm{P_{add}}, \mathrm{P_{EID}} \\ \mathrm{K_M} \leftarrow \mathrm{P_{name}}, \mathrm{P_{SSN}}, \mathrm{Gender} \end{cases} \tag{1}$$

Words with similar meanings have been given comparable numerical representations. The relevant terms in the CD have been found by the term frequency-inverse document frequency (TF-IDF). By using theTF-IDF technique the KE and simple text analysis have been performed. This TF-IDF method is unable to identify words that are semantic meaning-based. Thus, the semantic meaning-based data from the CD will be extracted by the Word2Vec algorithm. The SSD terms have been predicted using Algorithm 1. The values of TF-IDF have been calculated by using equations (2) and (3). The feature extraction (FE) extracts relevant data based on keywords, lemmas, and synonyms.

Algorithm 1. Information Extraction – Keyword and Sequence Extraction and Pattern Matching

```
Input: TD, Term Keyword List (TMKWL), Term Pattern (TMPAT)
Output: LT
Procedure: Term Identification (TD, TMKWL, TMPAT)
1. Initialize Term ⟵ 0
2. for all TextSeq, Seq. ∈ TD
      Matches each Pat∈ TMPATdo
      Term ⟵ <Seq.>
      end for
3. for all text Seq, Seq. ∈ TD
      Matches each Pat∈TMKWLdo
      Term ⟵ <Phrase><Seq.>
      end for
4. return Terms
```

$$\mathrm{TF(Term)} = \frac{\text{Number of Times Term Appears in a Document}}{\text{Total Number of Terms in the Document}} \tag{2}$$

$$\mathrm{IDF(Term)} = \log\left(\frac{\text{Total Number of Documents}}{\text{Number of Documents with Term}}\right) \tag{3}$$

Breaking down a data stream into a list of words, marking tokens related to parts of speech (POS), and associating useful lemma to POS, filtering the token list to obtain the most pertinent tokens, creating a list of features, and choosing the most pertinent features for the domain are performed by an above algorithm. Text tokenization, normalization, POS tagging, and lemmatization are the steps that make up FE. To DDig, these extracted traits are used.

The TF-IDF process extracts the general terms and not applicable for complex term extraction. In a word2vec algorithm, the CD is scanned entirely and the vector process is created for determining the often word identification. In this process, the semantic closeness between the words is identified by using the variables window, size and alpha. In word2vec, the unlabeled terms are identified by an artificial neural network. Table II shows the relevant term identification process of word2vec algorithm.Based on highest relevance, the terms are identified on the CD.

TABLE II. Word2Vec Similarity Identification – E.g Term – Great

Term	Relevance rate in percentage
Excellent	80
Fantastic	77
Perfect	74
Wonderful	70
Good	70
Amazing	63
Loved	62
Nice	62

3. Visual Feature Extraction

Medical images are used by doctors, radiologists, and pharmacists for drug discovery, DDig and treatment. The unique objects in the scanned images have been identified by VFE and extracted by scale-invariant FE technique. The convolution neural network (CNN) has been used for identifying abnormal object in medical images. The CNN extracts the features from an image's pixels and links them to the labels. The UMLS includes terms from the biomedical and health vocabularies. Hence, UMLS is used for identifying the extracted information. As a result, the visual features have been extracted from the images and then concepts are predicted from visual feature vector. The proposed NLP process is depicted in Fig. 2.

The CNN training complexity has been reduced by the 1 x 1024 length feature vector. With additional dimension reduction, CNN extracts highly relevant features. Two convolution layers (CL) and pooling layers, three dropout layers and one fully connected layer make up the proposed CNN network. The actual shape of the output is the same as the input using 64 and 128 filters, a four-size kernel, and a padding value of one for each of the two CL. Two max-pooling layers and one stride are used to acquire the most prominent features of the preceding feature map. The soft-max function and dropout layers address the over-fitting problems. Equation (4) describes how CNN output is expressed.

$$O_k^1 = f(s_k^1 + \sum_{i=1}^{N_{l-1}} VF_E(K_{ik}^{l-1}, t_i^{l-1})) \tag{4}$$

Where O_k^1 represents the output of kth neuron at the first layer, $f(\,)$ signifies the activation function, $VF_E(\,)$ represents the $VF_E(K_{ik}^{l-1}, t_i^{l-1})$

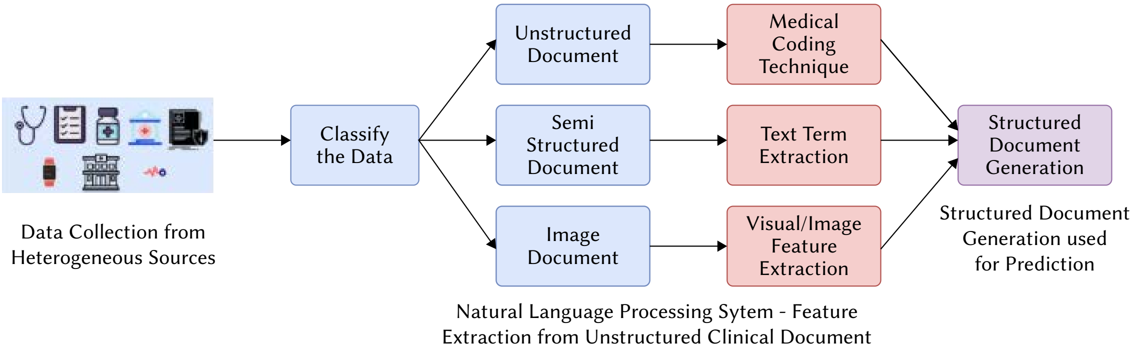

Classify the Data
Unstructured Document
Semi Structured Document
Image Document
Medical Coding Technique
Text Term Extraction
Visual/Image Feature Extraction
Structured Document Generation
Data Collection from Heterogeneous Sources
Structured Document Generation used for Prediction
Natural Language Processing Sytem - Feature Extraction from Unstructured Clinical Document

Fig. 2. NLP TEfrom USD CD.

denotes the kernel weight from the ith neuron at layer $l-1$ and its output. CNN encodes the data, and LSTM decodes it. The useful extracted terms are provided as an input to XAI to accurately predict diseases after being retrieved from USD, SSD and visual documents.

B. Explainable AI Based Disease Prediction

The prediction of the current AI is accurately predicting the results, but only little assistance given to the clinical expert (CE) in terms of DDig. Traditional methods for DDig include knowledge-based, data-driven, and hybrid approaches. These techniques fail to explain the disease forecasts. For the DM, high precision is required, and image specialists require much more information from the model than the binary prediction. ML may profit from advances in understanding that might lead to a clearer specification of its parameters in order to maintain impartiality in DM. ML accurately detects and explains the bias for training sets and tasks. The explanation can be obtained by describing which features are important for the output inference. Particularly in the medical industry, it is crucial that the interpretation of ML choices be paired with human interpretation. The previously trained ML model is used to predict and interpret new data. The evaluations of the layer, neuron, or prediction determine how to visualize the influence of pixels. The proposed XAI workflow is shown in Fig. 3.

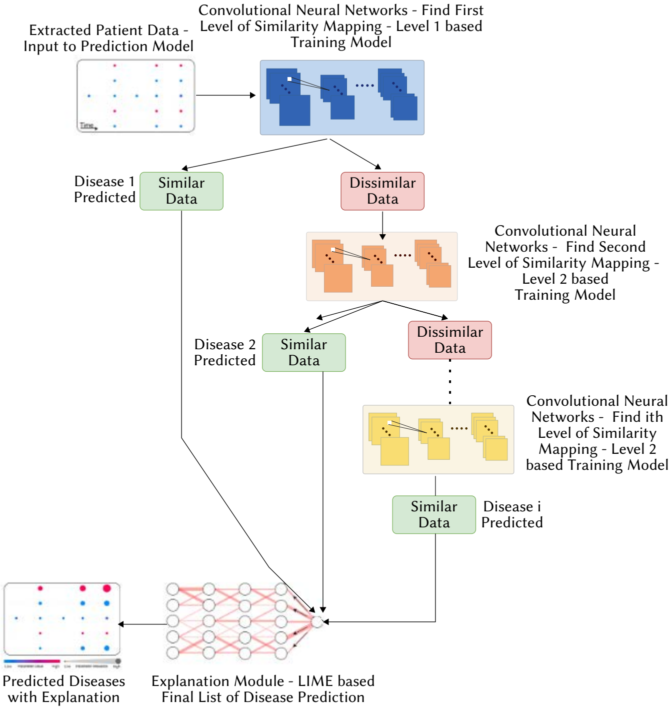

Extracted Patient Data - Input to Prediction Model
Convolutional Neural Networks - Find First Level of Similarity Mapping - Level 1 based Training Model
Disease 1 Predicted
Similar Data
Dissimilar Data
Convolutional Neural Networks - Find Second Level of Similarity Mapping - Level 2 based Training Model
Disease 2 Predicted
Similar Data
Dissimilar Data
Convolutional Neural Networks - Find ith Level of Similarity Mapping - Level 2 based Training Model
Similar Data
Disease i Predicted
Predicted Diseases with Explanation
Explanation Module - LIME based Final List of Disease Prediction

Fig. 3. Explainable AI Work Flow.

This phase took features from the NLP as input. On CNN, the first level similarity values have been predicted for identifying the "type 1 disease". The dissimilar data is transmitted to the next level of the CNN, and the level two mapped terms are classified from the dissimilar terms and labeled as "disease 2". With the use further disease data, the CNN next level has been trained and dissimilar values from the previous level is mapped to it. Certain terms have been mapped at each level, while others might not. This procedure continued, until all the retrieved terms have been mapped. Hence, this design is named as multi-level CNN model (CNNn). The CNNn prediction accuracy is high when compared to a single layer CNN model, and provides DDig information with sub-diseases. Only extracted features are examined for DDig. Hence, the minimal sized dissimilar terms are analyzed at each level. As a result, processing time of the proposed system is minimal. The XAI offers the prediction fairness, accessibility, interaction, causation, confidence, trustworthiness and privacy awareness. The explanations come in the form of text, images, local explanations, simplifications, and explanations that are relevant to the features. The LIME algorithm provides a justification based on expected outcomes. LIME analyzes the CNN output to determine how predictions alter as a result of various observations made during the subsequent procedure. The LIME technique's explanatory computation is performed using equation (5).

$$\delta(x) = argminL(f,g,\pi_x) + \omega(g), \qquad g \in G \tag{5}$$

Each model contained in the variable 'G' stands for decision trees, $\omega(g)$ denotes the number of trees and 'f' denotes the black box process. The variable 'L' is explained within the boundaries of the given locality π_x and the LIME approach aims to minimize 'L' and $\omega(g)$ to minimum value. Without any further requirement, the decision tree clearly self-explains the prediction.

C. Recommender System Based Disease Prediction

The RCS offers recommendations for products, services, and information to help users make better decisions. Based on user evaluations, the RCS determines recommendations. The RCS creates recommendations for new goods by applying a filtering algorithm to the input rating. Currently, RCS employs the Collaborative Filtering (CF) technique. Data sparsity and cold-start are two of the CF main drawbacks. These restrictions lead to the identification of bad recommendations. Hence, CF technique is not preferable to drug recommendation. Thus, the ACO RCS has been used to identify the drug recommendation in the proposed work. In an ACO, directed trust graph is utilized for finding the most reliable recommendations. The ACO selects the paths with the highest propagated trust values in order to determine the most effective suggestions. In order to properly manage and improve the accuracy of recommendations, the ant's updating approach is used in the proposed system. To raise the RCS accuracy rate, it is required to compute the semantic information

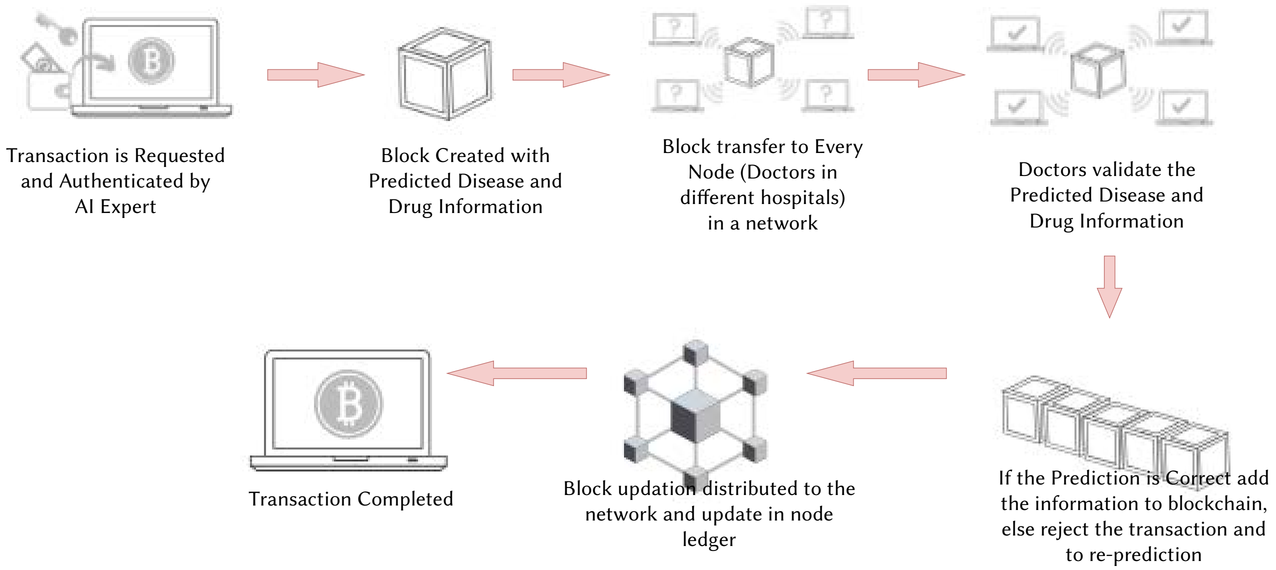

Transaction is Requested and Authenticated by AI Expert
Block Created with Predicted Disease and Drug Information
Block transfer to Every Node (Doctors in different hospitals) in a network
Doctors validate the Predicted Disease and Drug Information
If the Prediction is Correct add the information to blockchain, else reject the transaction and to re-prediction
Block updation distributed to the network and update in node ledger
Transaction Completed

Fig. 4. Block and Blockchain Generation.

and cluster the items according to their semantic similarity. The recommended drug for the disease is determined by semantic data and user ratings. AI experts Dig diseases and prescribe drugs, while clinical specialists validate the predictions. Equation (6) is used to measure the semantic similarity between two values.

$$Sim(X,Y) = \sum_{i=1}^{|V|} \left(\frac{Common(X,Y,V[i])}{Max(\deg(X,V[i]),\deg(Y,V[i]))}\right) * Weight(V[i]) \tag{6}$$

A set of data type properties and disease properties are contained in the vector V. Degree (A, V) specifies the number times a particular item X is related via property V, Common (X, Y, V) denotes the number of times item X and Y are associated with each other via property V, and weight (V) shows the significance of the property (V). The rating of all active drugs is determined by using equation (7) to determine the recommendation.

$$A_{u,i} = A_u + \frac{\sum_{v\in Nib_{u,i(t)}} (A_{v,i} - A_v) * IW^{c}_{uv}(t)}{\sum_{v\in Nib_{u\ i(t)}} IW^{c}_{uv}(t)} \tag{7}$$

Reflects the relative importance of the user v's rating in the context of 'c' at a time 't', where Au and Av are the mean rating of an active user 'u' and a neighbor 'v' for items in a certain group 'c' respectively. Based on similarity values and active users, the ACO Dig the diseases and drugs.

D. Blockchain Based Disease and Drug Information Transfer and Storage

Massive amounts of CD are currently stored in the cloud at lower maintenance and storage expenses. Numerous security vulnerabilities, including as the access, alternation, or deletion of CD, have happened in cloud storage as a result of CSP management. Serious problems arise as a result of these procedures in healthcare organizations. Therefore, a different storage technique is required to store patients' personal and medical information in secured and immutable way. BC is now essential to all industries, due to immutable storage of hash value (HV). The immutable storage provides lots of benefits to healthcare sector, such as to take an accurate decision, refer the drug history in future, etc. Hence, DDig and drug RCS are stored in BC is preferable. The BC network is established amongst clinical experts (CE) from various organizations. To decide the disease type and drug RCS, the BC network shares, data within CE.

If more than 50% of the CE accepts the prediction, the new block has been created and added to the BC network. This confirmation procedure, aids in increasing prediction accuracy and enhancing clinical patient care. The confirmed drug and disease information is saved in a block and added to BC network. The blocks are kept in a decentralized, distributed ledger makes it easier to analyze the past data in the future. The HV has been generated by using the secured hash algorithm 256 (SHA256). The SHA256 offers collision-free HV and its' simple to build blocks in a BC is compared to the other HV produced technique. Each block includes a timestamp (TS) value and details about the patient, disease, and drug. In order to analyze past data and track patient treatment information in the quickest time intervals, the TS is helpful. The BC workflow is shown in Fig. 4. Fig. 4 clearly shows the transaction initialization, CE validation, block and BC construction process. The HV of the block generation is generated by equation (8).

$$H_V(Patient_i) = SHA256(PB_{HV} + P_{ID} + Disease_{info} + Drug_{info}) \tag{8}$$

CE from various organizations validates the patient's information and responds to AI experts in the form of acceptance or rejection. If every prediction is approved, the relevant disease and drug information for the particular patient will be finalized. If not (rejected prediction), re-prediction is performed on the extracted data. Through this approach, DDig accuracy is highly improved and patients receive better clinical care.

VI. Experimental Results

A. Experimental Setup and Dataset Description

To evaluate the performance of the proposed system, our technique was implemented in 128GB RAM capacity system with python 3. The XAI was implemented with the help of Tensor-Flow library files. The private BC network was built by using Ganache based Ethereum environment. The IPFS version 0.4.19 was used for storing the blocks in off-chain storage location. In a ML performance analysis, the training and test data ratio was taken as 20:80. The BC network is created between patients, CE and AI experts. The patient node, able to view and access the DDig and recommended drug information. The CE node, able to verify the prediction and add blocks to BC network. The AI node, analyze the prediction and send the results to CE node.

The NLP terms based datasets had been collected from biomedical named entity recognition from https://github.com/cambridgeltl/MTL-Bioinformatics-2016 which consists of NCBI-disease, Bio-creative Vs. Chemical Disease Relation Extraction, BC2GM and JNLPAB. This dataset contains a large number of entities, sentences, words, training, development and test documents. The disease coding data has been taken from https://github.com/suamin/multilabel-classification-bert-icd10. Large-scale automated ICD coding from CD notes is included in this collection. The database at https://dailymed.nlm.nih.gov/dailymed/spl-resources-all-drug-labels.cfm which contains drug labeling summaries of information for safe and effective use of the drug, proposed by the manufacturer and approved by Food and Drug Administration, is used to generate the drug identification and labeling-based dataset. The chest x-ray image dataset is taken from IU collection for chest x-ray studies in Open-I (http://openi.nim.nih.gov/). This dataset includes a collection of x-ray images and related radiology reports are in a textual form.

B. Term Extraction Using NLP

The PM, KM, and SM processes have been used to carry out the TE. The TE process is depicted in Table III based on different types of extractions. By utilizing PM, KM, and SM, Date, age, email, phone, zip code, street and CD keywords have been extracted with a 91.47% accuracy of F1-measure.

TABLE III. Term Extraction Using PM, KM and SM

Type of Extraction	Precision (%)	Recall (%)	F1-measure (%)
PM	97.12	85.98	91.47
KM	93.06	81.98	87.18
SM	94.49	82.82	88.27

Table IV discusses the precision, recall, and F1-measure of the terms for patient ID, doctor name, zip code, CD number, city name, and hospital name.

TABLE IV. Term Extraction Using PM, KM and SM

Entity Name	Precision (%)	Recall (%)	F1-measure (%)
Patient ID	86.75	84.79	85.79
Doctor Name	96.60	83.80	89.75
Zip Code	100	94.39	97.10
CD number	96.10	91.75	93.84
City Name	83.96	78.56	81.22
Hospital Name	77.10	76.89	78.85

The combined approach of the PM, KM and SM provides higher prediction accuracy than the ML algorithm prediction technique. The ML algorithm prediction accuracy depends on the training dataset. Some expected level of false positive and negative values have been occurred due to inconsistent gold-standard.

C. Disease Prediction Using EAI

Precision, recall and F1-measure are the metrics used to assess the accuracy rate of CNNn-LSTM. The number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) is used to calculate these metrics. For the proposed system's training and testing, 60% of CT scans and X-ray images have been randomly chosen. The CNNn-LSTM parameters configuration is shown in Table V.

The sigmoid activation function, two fully connected layers, a batch size of 325, and 50 epochs, were used in the experiments. The training and testing losses for CNNn-LSTM accuracy have increased from 0.42 to 0.987 on the graph scale. From 1 to 5 epochs, the accuracy rate starts at 0.42 and rises to 0.987. The loss starts at 0.68 and gradually decreased until the 50th epoch. Table 6 compares the proposed CNNn-LSTM to DNN, GRU, RNN, LSTM, CNN1-LSTM, and CNN2-LSTM already in use.

TABLE V. CNNn-LSTM Parameter Configuration

Layer (Type)	Output Shape	Number of Parameters
Con_1	(None (N), 1023, 64)	256
M_pooling_1	(N, 512, 64)	64
Drop_out_1	(N, 512, 64)	64
LSTM_1	(N, 64)	131584
Droput_2	(N, 64)	64
Dense_Value_1	(N, 2)	258

TABLE VI. Performance Comparison of Proposed and Existing Techniques

Model Name	Precision (%)	Recall (%)	F1-Mesaure (%)	Accuracy (%)	Loss
DNN	91.78	87.96	86.78	87.64	1.78
GRU	94.89	94.87	94.85	94.86	0.13
RNN	94.69	94.23	93.96	93.95	0.10
LSTM	91.98	91.32	91.76	91.23	0.14
CNN1-LSTM	96.25	96.41	96.36	96.18	0.13
CNN2-LSTM	99.12	99.08	99.02	98.97	0.02
CNNn-LSTM	99.74	99.87	99.79	99.84	0.01

Till 24th epoch, the CNN1-LSTM model works well, after that, its performance suffers from the network's over-fitting. The CNN1-LSTM had an overall loss of 0.13. The CNN2-LSTM loss is below 0.02 and the proposed CNNn-LSTM loss is below 0.01. Based on various training and testing sets comparison, the suggested CNNn-LSTM outperforms than the CNN2-LSTM in terms of detection performance. Due to multiple levels in the proposed model, its accuracy rate is larger than other models.

D. Disease and Drug Recommendation Using ACO

In the medical industry, DDig's accuracy rate is essential. Drug recommendations are erroneous as a result of inaccurate DDig. The K-means, RF, and C4.5 algorithms as well as other popular clustering techniques are contrasted with the proposed ACO methodology. Fig. 5 illustrates a comparison of the proposed and current techniques. When compared to current techniques, the proposed ACO provides higher DDig accuracy and drug for each disease has now been identified.

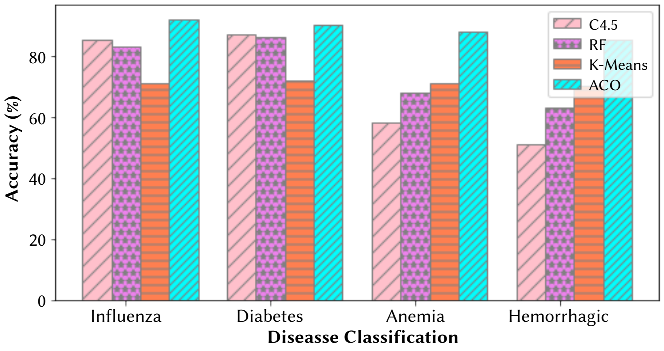


Fig. 5. Disease Classification Comparison.

TABLE VII. Drug RCS Analysis

Number of Patients	Ontology RCS				Proposed ACO RCS				Disease and Drug RCS
	Precision (%)	Recall (%)	Accuracy (%)	FMeasure(%)	Precision (%)	Recall (%)	Accuracy (%)	FMeasure(%)	
1	47.2	53.5	56.6	50.5	99.8	77.2	77.5	87.3	No
2	42.3	69.3	60.4	52.6	99.7	82.5	83.5	90.2	Yes, M
3	47.1	66.8	57.5	55.2	100	83.2	84.6	91.4	Yes, M
4	55.8	69.4	62.8	61.4	97.8	75.6	77.2	84.6	Yes, G
5	63.3	66.3	62.2	65.3	97.6	82.4	84.6	89.5	Yes, G
6	66.2	70.6	67.3	68.1	99.5	82.7	83.2	89.7	Yes, G
7	58.4	64.2	58.1	61.5	99.8	84.2	86.4	91.4	Yes, M
8	49.5	76.1	63.6	59.6	96.7	85.4	85.3	90.2	Yes, M
9	56.6	73.7	63.5	64.9	95.6	86.5	87.2	90.4	Yes, M
10	54.2	69.8	64.2	60.7	98.7	77.7	80.6	86.6	No
Average	54.4	67.2	61.4	59.5	97.8	81.5	83.7	89.4	-

TABLE VIII. DDigComparison - ML and BC Network

Parameters	LR	KNN	SVM	NB	DT	XGBoost	ANN	RF	Proposed BC
Precision (%)	74.58	81.36	78.65	65.24	80.75	82.54	82.53	84.98	99.56
Recall (%)	74.59	82.24	78.65	69.57	80.83	83.01	82.54	85.98	99.75
F1 Measure (%)	74.89	82.01	78.89	65.30	80.84	81.54	82.58	84.96	99.65
Accuracy (%)	74.25	82.05	78.65	62.78	80.84	82.45	82.58	84.79	99.58

For comparison, the diabetic drug RCS is analyzed in the proposed ACO and ontology-based RCS [30]. The results are demonstrated in Table VII. The results are diabetic [Yes, No] and drug [Metformin(M), Glindies (G)].

The accuracy rate increased from 61.4% to 83.7%, recall climbed from 67.2% to 81.5%, and F1-measure increased from 59.5% to 89.4% in the proposed method. The precision rate has been increased from 54.4% to 97.8%. By comparing the specific patient's prediction result with the average result, the information about the diabetes patient is precisely predicted for drug information. The drug is recommended to patients if their results are greater than average; otherwise, drug has not been recommended.

E. Disease and Drug Information Storage in B_c

The BC network receives the DDig and drug RCS from the previous stage. This information has been confirmed by more than 50% of the specialists on the BC network. After the DDig and drug RCS approval, the specialist adds the block to the BC network and notifies the requester. If not, the requester receives a message of refusal, and repeats the prediction based on other alternates. In table 8, the performance of the proposed BC network and ML algorithms like logistic regression (LR), k-nearest neighbor (KNN), support vector machine (SVM), naïve Bayes (NB), Decision tree (DT), artificial neural network (ANN), and random forest (RF) DDig performance have been compared.

The ratio of training and testing data for ML algorithms has been taken as 80:20. The proposed BC-based DDig predicts diseases with the highest accuracy rate when compared to ML algorithms. The accuracy rate of the proposed technique dependent on more than 50% of CE knowledge, not dependent on a single CE. Thus, the proposed technique's DDig prediction accuracy rate outperforms than ML algorithm's performance. Likewise, the drug RCS has been evaluated by the ML and BC network.

Now the block has been added to the BC network. The BC network's performance is affected by changes in the quantity of request or the patient-to-patient ratios. As a result, the performance of the BC networks has been analyzed by throughput, latency, and response time. The computational complexity of the CNNn-LSTM and ACO techniques are investigated in the following subsection.

VII. Performance Analysis

A. Computational Complexity

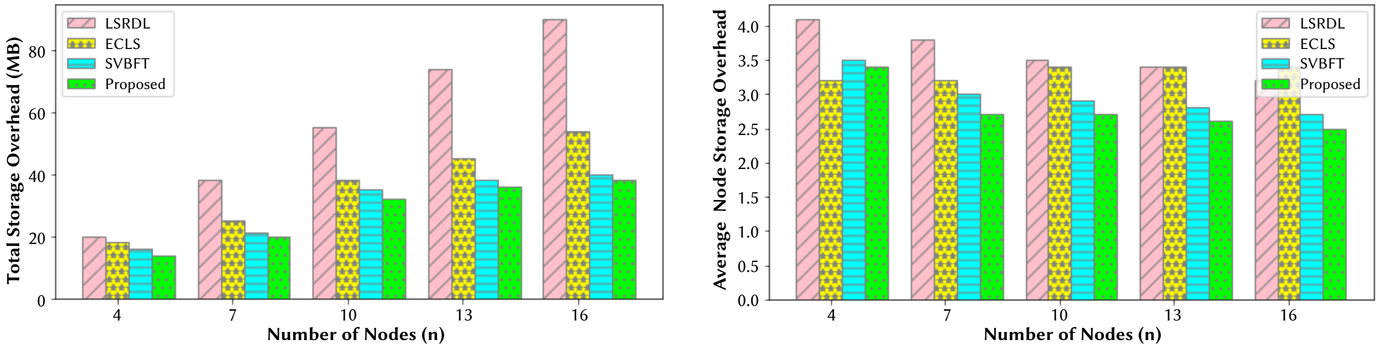
The CNNn-LSTM computational cost depends on the number of layers in CNNn. If $n_{c,1} = n/2$ and $n_{c,2} = n/4$ and the computational complexity of the CNN is $O(\lambda n^2)$ Here, λ denotes the size of the training set in sequential order. The complexity of the LSTM network is $O(4n_l\lambda(\frac{n^2 n_l}{16} + n_l)$ where nl is the number of neurons in each gate. Equation (9) computes the CNNn-LSTM network computational complexity of a single data sample.

$$O(\lambda n^2 \left(4n_l \left(\frac{n_l}{16} + n_l\right)\right) \quad (9)$$

Prediction complexity of the RCS for a 'n' active users and 'm' objects is O(n*m). The time required to compare the active users to the specific context of the other users is O(n*m). The ACO RCS has an O(nm+nm) computational complexity overall.

B. Storage Space Requirement Analytics

The limitation of BC is the block storage size. Each block can hold up to 1MB of data. The CD contains a variety of data types requires a lot of storage space because they are kept on the block in their original format. In order to securely store, the predicted results, recommended drug information and expert decisions are kept in the caretaker's locations. This storage technique requires less than 1MB of storage capacity for each patient. The data for every patient had been kept in a distinct block. The suggested system's storage overhead has been compared with conventional storage methods such as score-voting based Byzantine fault tolerance (SVBFT), erasure code-based low storage (ECLS) and low storage room distributed ledger (LSRDL). The comparison of the proposed and existing methodologies is shown in Fig. 6a and b [31]. The number of nodes has been regulated at 4 to 16 and the block size is set at 400 to 2800 transactions.



(a) (b)

Fig. 6. (a) Total storage Costs. (b) Node storage overhead on average.

As can be seen in fig. 6a and b, the proposed strategy has lower average and total overheads than existing techniques now in use. Under different storage techniques, the overall storage space requirement grows as the number of nodes does as well.

C. Storage B_cThroughput and Latency Time Analysis

Throughput (TP) Analysis: The average number of transactions that can be handled successfully per second (TPS) in the BC network is used to calculate TP. Total transactions divided by total time (time in seconds) have been used to calculate the BC network transaction time. For TP analysis, the total number of TPS has been used. The proposed work's TP analysis based on the operations such as opening the transaction, query processing and transfer data has been shown in Fig. 7.

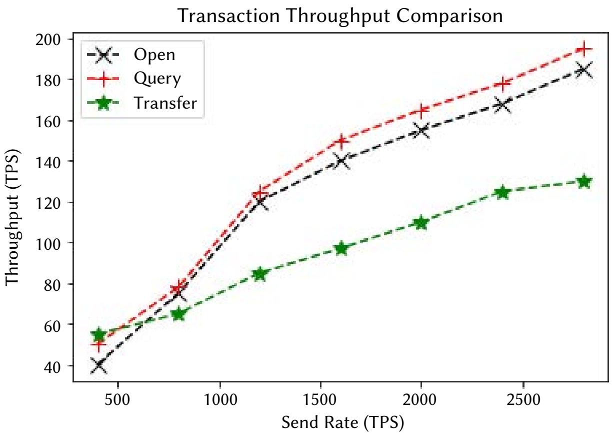


Fig. 7. Transaction Throughput Analysis.

Latency (LT) Analysis: The LT has been calculated from transaction submission confirmation. The propagation and processing times are included for determining the LT time. A processing rate of 400 to 2800 TPS has been used for measuring the BC network's performance, with 50 TPS has been taken into account for data opening, querying, and transferring. The LT begins when a transaction is submitted and ends when it is added in a BC network. With varying sending rates of 400, 800, 1200, 1600, 2000, and 2800 TPS, the average latency had been measured from 400 to 2800. The average LT rate increased along with the transmission rate. Fig. 8 shows the LT analysis of the proposed work.

VIII. Security Analysis

Data privacy for patient CD is offered by the proposed BC-based DDig and drug RCS storage technique. HV of CD, is stored in BC instead of actual data. For various CD sizes, an equal size of HV is produced as an output. This characteristic prevents data tracking of individuals. Instead of private storage, the CD is kept at the decentralized storage location. The decentralized storage technique avoids a single point of failure and authorized user access denying. The proposed technique differs from attribute based encryption (ABE) and key-aggregate cryptosystem (KAC) in terms of security support for CD. The proposed BC storage is compared with existing ABE and KAC techniques, as shown in table IX.

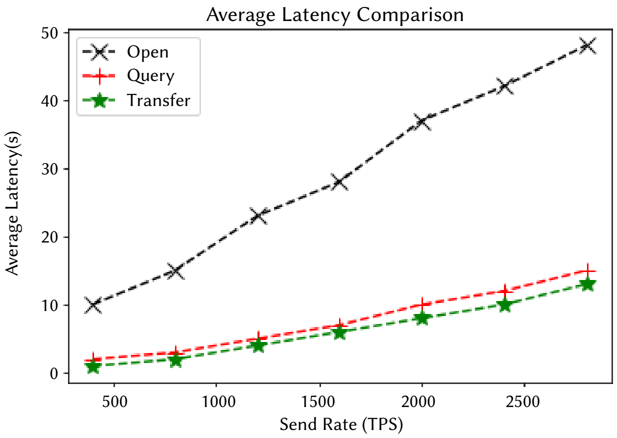


Fig. 8. Latency Time Analysis.

TABLE IX. Comparison of ABE, KAC and Proposed BC Technique

Security Parameter	ABE	KAC	Proposed BC
Tamper Resistance	√	√	√
Information Traceability	√	√	X
Secure Storage	√	√	√
Privacy Protection	√	√	√
Access Log	X	X	√
Access Period	X	X	√
Reliance on Trusted Third Parties	√	√	X
User Anonymity	X	X	√
Control of CD	Incomplete	Incomplete	Complete

Attack analysis of proposed technique:

- Sybil Attack: In a proposed work, the BC network was created between the registered and authorized users in a permissioned network. The attackers are unable to participate in the conversion and access the information. Hence, unauthorized access is completely avoided with this technique.

- Spoofing attack: The authorized user devices are synchronized with the network services. Hence, the authorized users are able to access the data through the registered devices. If an unregistered device is trying to enter into the network, the alter message is sent to the caretakers. Hence, the spoofing attacks are impossible in this proposed technique.
- Man-in-the-middle attack: the information is stored in the form of HV. If the attacker tries to access information at the time of data transfer, they are unable to identify the information. Hence, middle-in-the-middle attack is not possible.
- Denial of service attack: In a proposed network, the registered users can only initiate the transaction and transfer the information. Other users are unable to enter into the network. Hence, the denial of service attack is eliminated in the proposed technique.

IX. Conclusion

To facilitate easy maintenance, analysis, and transfer, the health information is kept in digital form. The knowledge, accessibility, and experience of the professionals determine how well manual disease diagnosis and drug prediction perform in the medical area. The specialist's availability has a significant impact on the patient's quality of life. The artificial intelligence helps in making decisions that are accurate but does not reveal decision details. The specific hospital determines whether it can handle emergency patients. The clinical data is in the form of structured and unstructured forms like numerical results, text prescriptions, scanned images, and other types. It is important to manage unstructured data carefully while making decisions. With more than 91% accuracy, a natural language processing system analyses the unstructured data to distinguish between distinct data types, and the CNNn-LSTM technique accurately diagnosed diseases with higher than 99% accuracy. An ant colony optimization-based recommender system analyses the prediction and more precise drug for the disease. The drug decision and disease information have been saved in a permissioned blockchain after expert approval. More than 50% of clinical experts validate the decisions. As a result, judgments are made in emergency situations as quickly and precisely as feasible. The computational complexity and storage space requirements of the proposed system are lower than those of the existing techniques. Similarly, several aspects of blockchain throughput, latency, and security analysis have been examined. In a proposed work, the patient and clinical experts are having complete control over the clinical data. In future, the proposed technology will be extended for dynamic data collection in remote locations and decision-making.

References

[1] Gazali, S., S. Kaur, and I. Singh, "Artificial intelligence based clinical data management systems: A review," *Informatics in Medicine Unlocked*, vol. 9, pp. 219–229, 2017.

[2] A. Martinez-Millana, A. Saez-Saez, R. Tornero-Costa, and N. Azzopardi-Muscat, "Artificial intelligence and its impact on the domains of universal health coverage, health emergencies and health promotion: An overview of systematic reviews," *International Journal of Medical Informatics*, vol. 168, pp. 1–12, 2022.

[3] S. M. Sumathi, N. Vijayaraj, S. P. Raja, and M. Rajakamal, "Internet of things based confidential healthcare data storage, access control and monitoring using blockchain technique," *Computing and Informatics*, vol. 41, no. 5, pp. 1–31, 2022.

[4] S. M. Sumathi, S. Sangeetha, and A. Thomas, "Generic cost optimization and secured sensitive attribute storage model for template-based text document on cloud," *Computer Communications*, vol. 150, pp. 569–580, 2020.

[5] V. Bellini, P. Pelosi, M. Valente, A. V. Gaddi, and M. Baciarello, "Using artificial intelligence technique to support clinical decisions in perioperative medicine," *Perioperative Care and Operating Room Management*, vol. 28, pp. 1–8, 2022.

[6] J. R. Parikh, C. A. Genetti, A. Aykanat, C. A. Brownstein, and K. Schmitz-Abe, "A data-driven architecture using natural language processing to improve phenol-typing efficiency and accelerate genetic diagnoses of rare disorders," *HGG Advances*, vol. 2, no. 1, pp. 1–10, 2021.

[7] J. Lve, N. Viani, J. Kam, L. Yin, S. Verma, S. Puntis, R. N. Cardinal, A. Roberts, R. Stewart, and S. Velupillai, "Generation and evaluation of artificial mental health records for natural language processing," *Digital Medicine*, vol. 3, no. 1, pp. 1–10, 2020.

[8] E. H. Houssein, R. E. Mohamed, and A. A. Ali, "Machine learning techniques for biomedical natural language processing: A comprehensive review," *IEEE Access*, vol. 9, pp. 1–26, 2021.

[9] C. Comito, D. Falcone, and A. Forestiero, "AI-driven clinical decision support: Enhancing disease diagnosis exploiting patient similarity," *IEEE Access*, vol. 10, pp. 1–12, 2022.

[10] S. Sim and M. Cho, "Convergence model of AI and IoT for virus disease control system," *Personal and Ubiquitous Computing*, vol. 25, pp. 1–11, 2021.

[11] H. Lu, S. Uddin, F. Hajati, M. A. Moni, and M. Khushi, "A patient network-based machine learning model for disease prediction: The case of type-2 diabetes mellitus," *Applied Intelligence*, vol. 51, pp. 4667–4680, 2021.

[12] F. Amato, L. Coppolino, G. Cozzolino, G. Mazzeo, F. Moscato, and R. Nardone, "Enhancing random forest classification with NLP in DAMEH: A system for data management in e-health domain," *Neurocomputing*, vol. 457, pp. 79–91, 2021.

[13] E. L. Romm and I. F. Tsigelny, "Artificial intelligence in drug treatment," *Annual Review of Pharmacology and Toxicology*, vol. 60, pp. 353–369, 2020.

[14] E. Khodabandehloo, D. Riboni, and A. Alimohammdi, "HealthXAI: Collaborative and explainable AI for supporting early diagnosis of cognitive decline," *Future Generation Computer Systems*, vol. 125, pp. 168–189, 2021.

[15] T. Ploug and S. Holm, "The four dimensions of contestable AI diagnostics: A patient-centric approach to explainable AI," *Artificial Intelligence in Medicine*, vol. 107, pp. 1–5, 2020.

[16] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM: An effective heart disease prediction model for a clinical decision support system," *IEEE Access*, vol. 8, pp. 115710–115726, 2020.

[17] R. F. Mansour, A. E. Amraoui, I. Nouaouri, V. G. Diaz, D. Gupta, and S. Kumar, "Artificial intelligence and internet of things enabled disease diagnosis model for smart healthcare systems," *IEEE Access*, vol. 9, pp. 163512–163521, 2021.

[18] V. Singh and D. Jain, "A hybrid parallel classification model for the diagnosis of chronic kidney disease," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, pp. 8–15, 2021.

[19] S. Dinakaran and P. Anitha, "An efficient drug compound analysis using spectral deep feature classification based compound analysis model for drug recommendation," *Neuroscience Informatics*, vol. 2, no. 3, pp. 1–9, 2022.

[20] F. Ali, S. El-Sappagh, S. M. R. Islam, and A. Ali, "An intelligent healthcare monitoring framework using wearable sensors and social networking data," *Future Generation Computer Systems*, vol. 114, pp. 23–43, 2021.

[21] F. Ali, S. M. R. Islam, D. Kwak, P. Khan, and N. Ullah, "Type-2 fuzzy ontology-aided recommendation systems for IoT-based healthcare," *Computer Communications*, vol. 119, pp. 138–155, 2018.

[22] L. F. Grandamorales, P. Valdiviezo-Diaz, R. Reategui, and L. Barba-Guaman, "Drug recommendation system for diabetes using a collaborative filtering and clustering approach: Development and performance evaluation," *Journal of Medical Internet Research*, vol. 24, 2022.

[23] Q. Ye, C.-Y. Hsieh, Z. Yang, Y. Kang, J. Chen, D. Cao, S. He, and T. Hou, "A unified drug-target interaction prediction framework based on knowledge graph and recommendation system," *Nature Communications*, vol. 12, no. 1, pp. 1–12, 2021.

[24] G. Zhang, X. Chen, L. Zhang, B. Feng, X. Guo, J. Liang, and Y. Zhang, "STAIBT: Blockchain and CP-ABE empowered secure and trusted agricultural IoT blockchain terminal," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 5, pp. 56–65, doi: 10.9781/ijimai.2021.12.002.

[25] T. Hovorushchenko, A. Moskalenko, and V. Osyadlyi, "Methods of

medical data management based on blockchain technologies," *Journal of Reliable Intelligent Environments*, vol. 9, no. 1, pp. 5–16, 2022.

[26] S. Singh, S. K. Sharma, P. Mehrotra, P. Bhatt, and M. Kaurav, "Blockchain technology for efficient data management in healthcare system: Opportunity, challenges and future perspectives," *Materials Today: Proceedings*, vol. 62, pp. 5042–5046, 2022.

[27] R. Cerchione, P. Centobelli, E. Riccio, S. Abbate, and E. Oropallo, "Blockchains coming to hospital to digitalize healthcare services: Designing a distributed electronic health record ecosystem," *Technovation*, vol. 117, pp. 1–16, 2022.

[28] G. Lin, H. Wang, J. Wan, L. Zhang, and J. Huang, "A blockchain-based fine-grained data sharing scheme for e-healthcare system," *Journal of Systems Architecture*, vol. 127, 2022.

[29] K. Johnson, "Implementation of ICD-10: Experiences and lessons learned from a Canadian hospital," in *IFHRO Congress, AHIMA Convention*, Oct. 2004. [Online]. Available: www.ahima.org.

[30] D. Riano, F. Real, J. A. López-Vallverdu, F. Campana, S. Ercolani, P. Mecocci, R. Annicchiarico, and C. Caltagirone, "An ontology-based personalization of health-care knowledge to support clinical decisions for chronically ill patients," *Journal of Biomedical Informatics*, vol. 45, no. 3, pp. 429–446, 2012.

[31] C. Li, J. Zhang, X. Yang, and Y. Luo, "Lightweight blockchain consensus mechanism and storage optimization for resource-constrained IoT devices," *Information Processing and Management*, vol. 58, no. 6, pp. 1–24, 2021.

Sumathi M.

Sumathi M. completed her B. Eng. in computer science and engineering in 2003 from the Shri Angalamman College of Engineering and Technology, Tiruchirappalli. She completed her M.Tech. in information technology in 2008 from the Bharathidasan University, Tiruchirappalli. She completed her Ph.D. in 2021 in the area of data security from the National Institute of Technology, Tiruchirappalli. Currently, she is working as Assistant Professor in the School of Computing at SASTRA Deemed University, Thanjavur, Tamil Nadu, India.

S. P. Raja

S. P. Raja is born in Sathankulam, Tuticorin District, Tamilnadu, India. He completed his schooling in Sacred Heart Higher Secondary School, Sathankulam, Tuticorin, Tamilnadu, India. He completed his B. Tech in Information Technology in the year 2007 from Dr. Sivanthi Aditanar College of Engineering, Tiruchendur. He completed his M.E. in Computer Science and Engineering in the year 2010 from Manonmaniam Sundaranar University, Tirunelveli. He completed his Ph.D. in the year 2016 in the area of Image processing from Manonmaniam Sundaranar University, Tirunelveli. Currently he is working as an Associate Professor in the School of Computer Science and Engineering in Vellore Institute of Technology, Vellore, Tamilnadu, India.

Trends in Addiction to Psychoactive Substances Among Homeless People in Colombia Using Artificial Intelligence

Hugo Ordoñez[1]*, Ricardo Timarán-Pereira[2], Juan-Sebastián González-Sanabria[3]*

[1] Universidad del Cauca, Popayán (Colombia)
[2] Universidad de Nariño, Pasto (Colombia)
[3] Universidad Pedagógica y Tecnológica de Colombia, Tunja (Colombia)

* Corresponding author: juansebastian.gonzalez@uptc.edu.co (J. S. González-Sanabria), hugoordonez@unicauca.edu.co (H. Ordoñez).

Received 31 July 2023 | Accepted 22 November 2023 | Published 5 February 2024

Abstract

Introduction: Currently, homelessness should not be seen as just another problem, but as a reality of inequality and the absence of social justice. In this sense, homeless people are subjected to social disengagement, lack of job opportunities or the instability of these, insecurity circumstances, these aspects being one of the causes associated with the consumption or addiction to psychoactive substances. *Data*: To define the proposed approach, data from the Census of Street Inhabitants - CHC- 2021 of the National Administrative Department of Statistics (DANE), which contains 19,375 records and 25 columns, were used. *Methodology*: This article presents an artificial intelligence approach that implements a model based on machine learning algorithms for identifying addiction trends to psychoactive substances in street dwellers in Colombia. *Conclusions*: Based on the results obtained, it is evident that the approach can serve as a support for decision making by municipal administrations in the definition of social public policies for the street-dwelling population in Colombia.

Keywords

Addiction, Artificial Intelligence, Homelessness, Psychoactive Substances, Social Justice.

DOI: 10.9781/ijimai.2024.02.002

I. Introduction

In today's world, homelessness should not be seen as just another problem but as a reality of inequality and lack of social justice. In this sense, it is common to observe citizens in large cities who often transit or live permanently on the streets: children, young people, adults, older people, and even families, who, regardless of their age, sex, race, marital status, social condition, mental condition, or occupation, live there permanently or for prolonged periods, making life in this context a transitory or long-lasting option [1].

The homeless population has increased due to political, economic, and cultural factors that affect social organization. Among these, we can name displacement, armed conflict, domestic violence, and unemployment. As a particular case and focus of attention and social degradation, the consumption of psychoactive substances is identified as the most substantial and explosive factor that generates the phenomenon of homelessness [2]-[3].

Homeless people are subject to social disengagement, lack of employment opportunities or instability in this area, and circumstances of insecurity. Additionally, the experiences of loss, abandonment, and domestic violence are some causes associated with illicit psychoactive substance use or addiction [4]. The problem at the beginning of the addiction of these people is strongly related to the rupture of meaningful or affective bonds, family issues, and hostile environments. Similarly, emotional abandonment, neglect or permissiveness on the part of the caregiver or parents increases the risk of addiction [5].

Moreover, addiction arises due to the properties of illicit psychoactive substances, which, by acting more rapidly within the organism and being eliminated more swiftly, foster an increased compulsion for the individual to acquire them to sustain their consumption. The street is the most accessible place to acquire these substances [6].

Currently, as an alternative to face transcendental problems in society, technological tools have allowed to establish new state or governmental policies based on data evidence. For example, artificial intelligence tools, specifically machine learning (ML) algorithms, are being used to modernize services and assist governments in their decision-making regarding social policy issues [7]-[11].

Therefore, this paper presents an ML model for identifying trends of addictions to illicit psychoactive substances in homeless people

Please cite this article in press as:
H. Ordoñez, R. Timarán-Pereira, J. S. González-Sanabria. Trends in Addiction to Psychoactive Substances Among Homeless People in Colombia Using Artificial Intelligence, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 52-60, 2025, http://dx.doi.org/10.9781/ijimai.2024.02.002

in Colombia. The data of the Homeless People Census (CHC) from 2021 of the National Administrative Statistics Department (DANE), containing 19,375 records and 25 columns, was used to define the proposed model.

The model evaluation results allow identifying the trends of addictions to psychoactive substances in relation to the age of each homeless person. Moreover, the results can help in the decision-making of municipal administrations to define social public policies for the care of these people in the country.

This paper is structured into five sections. Section one describes the work related to the subject of the study. Section two presents the motivation context. Section three explains the proposed model. Section four presents the evaluation. Finally, section five develops the conclusions and future work.

II. Motivation and Related Works

Consumption, abuse, and addiction to psychoactive substances, licit or illicit, are a matter of public health. Thus, additional efforts must be made to evaluate this phenomenon in homeless people since the lifestyle of this group is related to higher drug dependence. Several studies have addressed this topic.

One study [12] addresses the control of anxiety in cocaine users undergoing outpatient addiction treatment. It applies six steps of the intervention mapping approach: needs evaluation, creation of performance objective matrices, method selection and practical strategies, program development, adoption and implementation, and evaluation, to develop the interpersonal nursing theory for anxiety in the intervention of persons with illicit psychoactive substance use disorders.

Similarly, other studies [13]-[14] address some differences between women who are mothers and those who do not have children in a sample of women living on the street in Madrid, Spain. The information was collected through a structured interview. The results evidence that the women living in the street were mothers who had experienced traumatic situations from an early age and had higher levels of illicit psychoactive substance abuse. Furthermore, they had issues with the judiciary system, which may have negatively impacted the lives of their children and their relationships.

Likewise, another research [15] aimed at identifying the structure of the social representations homeless people have of persons in the same situation who consume drugs. This was based on the social representations theory and addressed 158 homeless people in the historic center of Salvador, Bahía, Brazil. The data were collected through free word association using the inductor homeless people who consume drugs. The data were analyzed with two software programs. The results identified that the participants were mainly young Afro-descendant men who had finished primary school. It was concluded that these people have their lives at risk, are excluded, and need help.

Another study [16] focused on analyzing the knowledge and experiences with new psychoactive substances among users in the homeless population. The participants were selected from support charitable organizations in the United Kingdom through convenience sampling. Descriptive and logistic regression statistics were applied to analyze the obtained data. The results showed that the street dwellers consumed illicit psychoactive substances to escape reality and self-medicate, and they stopped consuming due to the adverse side effects. These effects were reported by most of the participants and caused over 20% of them to require medical treatment.

In the same sense, some studies [17]-[18] examined aspects such as sociodemographic characteristics, access to economic resources, social support, addiction chronicity and access to new technologies of the homeless people in León, Nicaragua. A questionnaire was used to collect the data. The results showed that homeless people have social difficulties and high chronicity levels. Despite the major cultural and developmental differences between Spain and Nicaragua, there are significant similarities among the homeless people in both countries.

There have also been studies in Colombia regarding homeless people. One of them [19] described their health situation, concluding that one in five people has a health issue, such as dental problems, respiratory problems, abdominal pain or injuries caused by third parties. The most common chronic diseases are hypertension and diabetes.

Similarly, another research [20] described that homelessness affects health and makes these people vulnerable. The study was based on a program called Mobile Care Center for Drug-Dependent People (CAMAD). Its objective was to interpret the experiences of a group of homeless people in the Rafael Uribe area in Bogotá, Colombia, focusing on health.

In conclusion, several studies and research on homeless people worldwide address issues related to their addiction to illicit psychoactive substances and their health. Although these studies have had outstanding results, none have taken advantage of the potential of current technologies, specifically artificial intelligence, to analyze the data collected. Therefore, the ML model presented in this paper analyzes socioeconomic, health, economic activity, gender, and types of addiction variables to predict how these addictions may be related to the age of street dwellers in Colombia.

In Colombia and most of the world, the homeless person is subjected to social disengagement, job instability, and precarious conditions. The pressure of other homeless people or friends, the emotional suffering, anxiety, depression, and the environmental stress experienced on a daily basis when living on the street can be factors that increase the risk of drug use [21].

In Colombia, there are 22,790 homeless people; 12% are women, and 88% are men [22]. Fig. 1 shows the number of homeless people per department. The departments of Valle del Cauca and Antioquia are noteworthy because their capitals, Cali and Medellín, respectively, have a multicultural population. Additionally, they receive many migrants from the center of the country due to their increased industrialization and development.

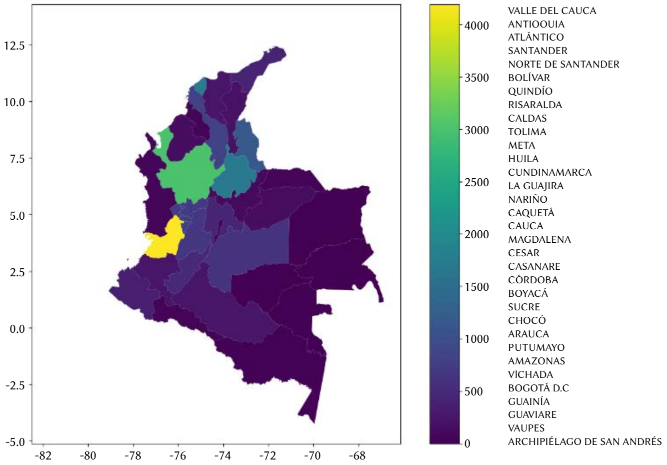


Fig. 1. Population of homeless people per department in Colombia.

Illicit psychoactive substances contain natural or synthetic compounds that act on the nervous system generating alterations in the functions that regulate thoughts, emotions, and behavior. These substances are not freely distributed and are punishable by law [23].

Some of them are:

- Marijuana or cannabis: It directly affects brain function, particularly the brain parts responsible for memory, learning, attention, decision-making, coordination, emotions, reaction time, relaxation and euphoria, anxiety, fear, distrust, and panic.
- Basuco: It is a toxic substance whose main risks when consumed are related to neurological and physical deterioration. It destroys brain tissue and causes irreversible memory loss; the effects are immediate: the skin turns yellow, the lips get dried, the tongue gets numb, the pupils dilate, and the body trembles. Its dissolution in the bloodstream is swift, which makes it very addictive.
- Cocaine: It affects the nervous system and the rest of the body immediately. These affectations include vasoconstriction, mydriasis, hyperthermia, tachycardia, and hypertension. Additionally, the effects derived from the euphoria, mainly during the first 30 minutes, are hyperstimulation, the sensation of less tiredness, and a state of greater mental alertness.
- Heroin: It is highly addictive, and its effects are very pleasant, which causes a continuous and repetitive consumption behavior. Heroin gets to the brain rapidly and adheres to the opioid receptors of the cells, especially those associated with pain and pleasure and those controlling the heart rate, sleep, and breathing. It produces dryness in the mouth, reddening and heating of the skin, heaviness in arms and legs, nausea and vomiting, intense itching and clouding of the mental faculties.

Psychoactive substance consumption is a problem affecting society in general, regardless of age, culture or social status. The consumption pattern of these substances in homeless people depends mainly on the age and type of substance. Fig. 2 presents the number of homeless people using each drug. It can be noted that the age range of basuco consumption is between 16 and 64 years, with a concentration in people between 25 and 50 years. Basuco is the most consumed substance among homeless people. Its characteristics increase addiction because its effects appear and vanish rapidly, making the person feel a greater need for consumption [24].

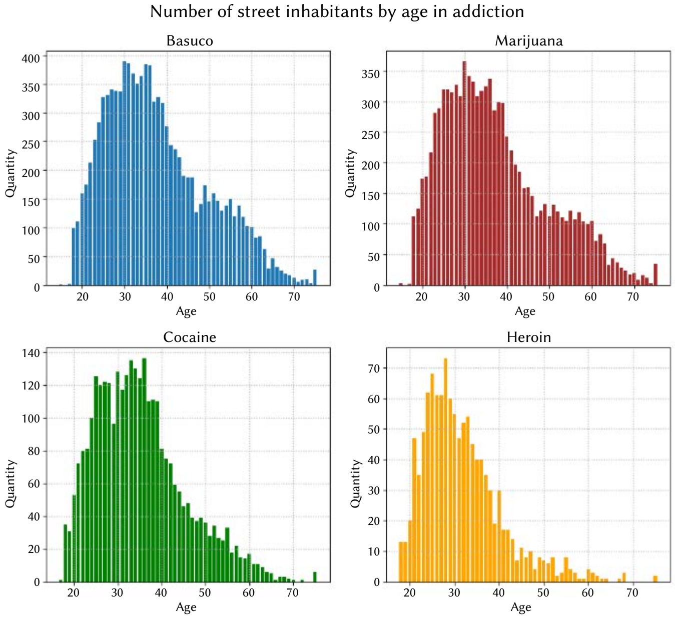


Fig. 2. Number of consumers per age according to the type of psychoactive substance.

Marijuana is the second most consumed substance by homeless people due to its effects: relaxation, drowsiness, a sensation of slowness in the passage of time, disinhibition, excessive joy, and eye reddening. The consumption age range for this drug goes from 19 to 60 years. Basuco and marijuana are the cheapest and easiest substances to acquire on the street [25].

Cocaine and heroin are consumed by people in similar age ranges (18 and 60 years old), but fewer people are addicted to these substances due to their price, the difficulty of acquiring them, and the symptoms they generate, such as restlessness, irritability, and anxiety. They can also cause tremors, dizziness, muscle spasms or paranoia, and serious medical complications if consumed excessively [26].

Addiction to illicit psychoactive substances becomes a disease that disturbs the brain and behavior of the person who consumes them and results in an inability to control consumption. Then, despite the damage they cause, it is customary to continue consuming them. Addiction to multiple substances is common among homeless people. Furthermore, as consumption increases, it is increasingly difficult to live without them [27].

Fig. 3 shows addictions to multiple illicit psychoactive substances in homeless people according to their age. Most notably, the substances of greatest consumption are marijuana and basuco, which means that the person who is addicted to marijuana is also addicted to basuco, and this preference oscillates among people between the ages of 18 and 60 years old.

Likewise, and more concerning, a smaller number of people are addicted to marijuana and cocaine. The age range of this group is 19 to 55 years old, and it is possibly related to the higher cost of cocaine compared to that of marijuana and basuco.

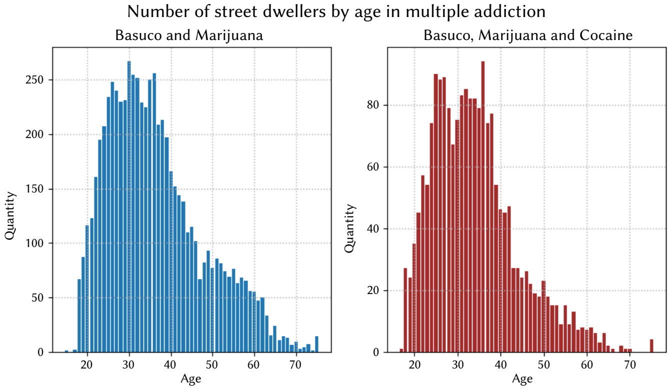


Fig. 3. Number of homeless people with multiple addictions.

Another relevant characteristic of homeless people is their source of income. As seen in Fig. 4, one of the primary sources of income is recycling, which is very common in major cities. Similarly, cleaning car windows and car guarding are ways of acquiring money and being self-employed for these people. Begging is also a widespread activity. Although not forbidden, it is considered a social problem directly related to inequality and poverty. It has to be noted that not all homeless people resort to theft to make ends meet, i.e., not all of them are criminals, only a small percentage resort to this activity.

People living on the street are exposed to malnutrition, have health problems, and are at risk of getting sick, dying, or experiencing diverse and continuous violent aggressions, which affect their health, due to the environment where they live [28]. Fig. 5 presents the most common aggressions suffered by homeless people. As it can be seen, the greatest fear is losing their life violently, followed by beatings (to which they are exposed daily), and assaults with cold weapons, which are very common in this context.

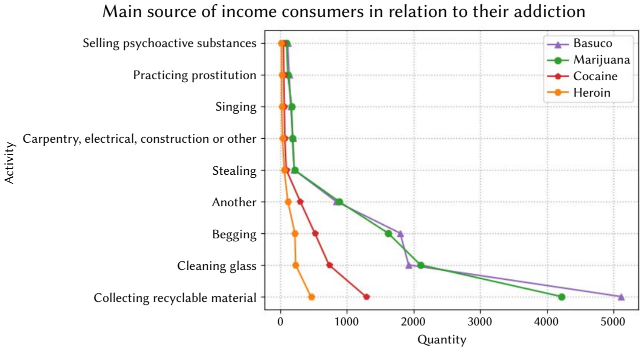


Fig. 4. Source of income of homeless people according to their addiction.

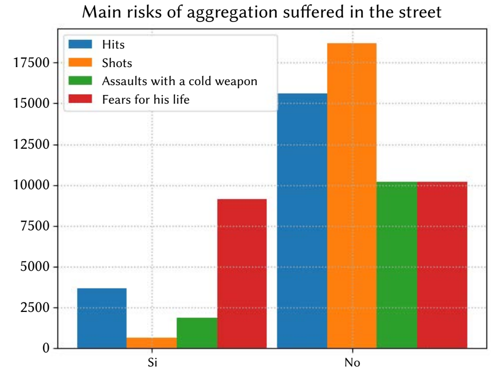


Fig. 5. Risks and fears of homeless people.

A high percentage of homeless people have addictions or are direct consumers of illicit psychoactive substances, and they are exposed to high health-related risks. Many of them are frequent victims of several crimes, and some of them commit crimes on a regular basis. The homeless person who consumes drugs lacks a source of employment, is concerned about having a permanent source of drug supply, uses drugs to "remedy" negative feelings, withdraws from friends and family, may trade friends for people who are regular users, and experiences increased tolerance and ability to process the drug. Thus, using technology, such as artificial intelligence, is relevant to support governmental entities in decision-making processes. Particularly when discussing programs that implement social strategies, such as the Comprehensive Policy for the Prevention and Care of Psychoactive Substance Use [29] and the Social Public Policy for Street People 2021-2032 [30].

III. Proposed Model

The predictive model for addiction to illicit psychoactive substances of the homeless people in Colombia is represented in Fig. 6. It applies a series of machine learning algorithms that integrate specific characteristics of a training dataset to make predictions and find trends. It is divided into three parts: data and pre-processing, machine learning models, and evaluation.

A. Data and Pre-processing

The data of the CHC from 2021 of the DANE containing 19,375 records and 25 columns were used to define the proposed model. The dataset related the information with the volume and the main socioeconomic and demographic characteristics of the homeless people in Colombia. It was built with comma-separated values (CSV) files containing information of the CHC from several years. A relational database was built based on these files, assimilating all the information. Subsequently, a minable view was generated through SQL statements to run the model. Table I presents the variables and their description.

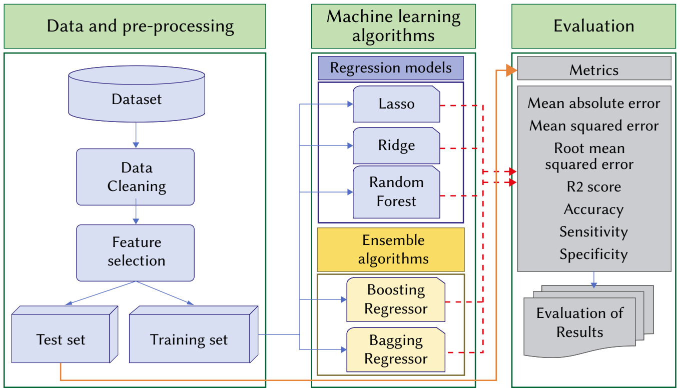


Fig. 6. Prediction model of addiction to illicit psychoactive substances based on artificial intelligence.

TABLE I. Variables of the Dataset

Variable	Description
Age	Age at the moment of the census
Gender	Gender of the homeless person
place_where_they_sleep	The place where the person usually sleeps or stays at night
hypertension	If the person has health problems related to hypertension
diabetes	If the person has health problems related to diabetes
Cancer	If the person has health problems related to cancer
tuberculosis	If the person has health problems related to tuberculosis
hiv_aids	If the person has health problems related to AIDS
reason_for_living_on_the_streets	What is the main reason for living on the streets
time_living_on_the_streets	How long the person has been living on the streets
reason_for_staying_on_the_streets	Reason for staying on the streets, what forces the person to stay on the streets
source_of_income	What is the person's source of income, how do they access money to live
Tobacco	Uses or smokes tobacco
Alcohol	Uses or is addicted to alcohol
marijuana	Uses or is addicted to marijuana
inhalants	Uses or is addicted to inhalants
Cocaine	Uses or is addicted to cocaine
Basuco	Uses or is addicted to basuco
Heroin	Uses or is addicted to heroin
Pills	Uses or is addicted to pills
other_drugs	Other drugs that the person uses frequently
fear_for_their_life	Fears for their life while living on the streets
beating_victim	The person has been a victim of beatings while living on the streets
gunshot_victim	The person has been a victim of gunshots while living on the streets
cold_weapon_victim	The person has been a victim of cold weapon assaults while living on the streets

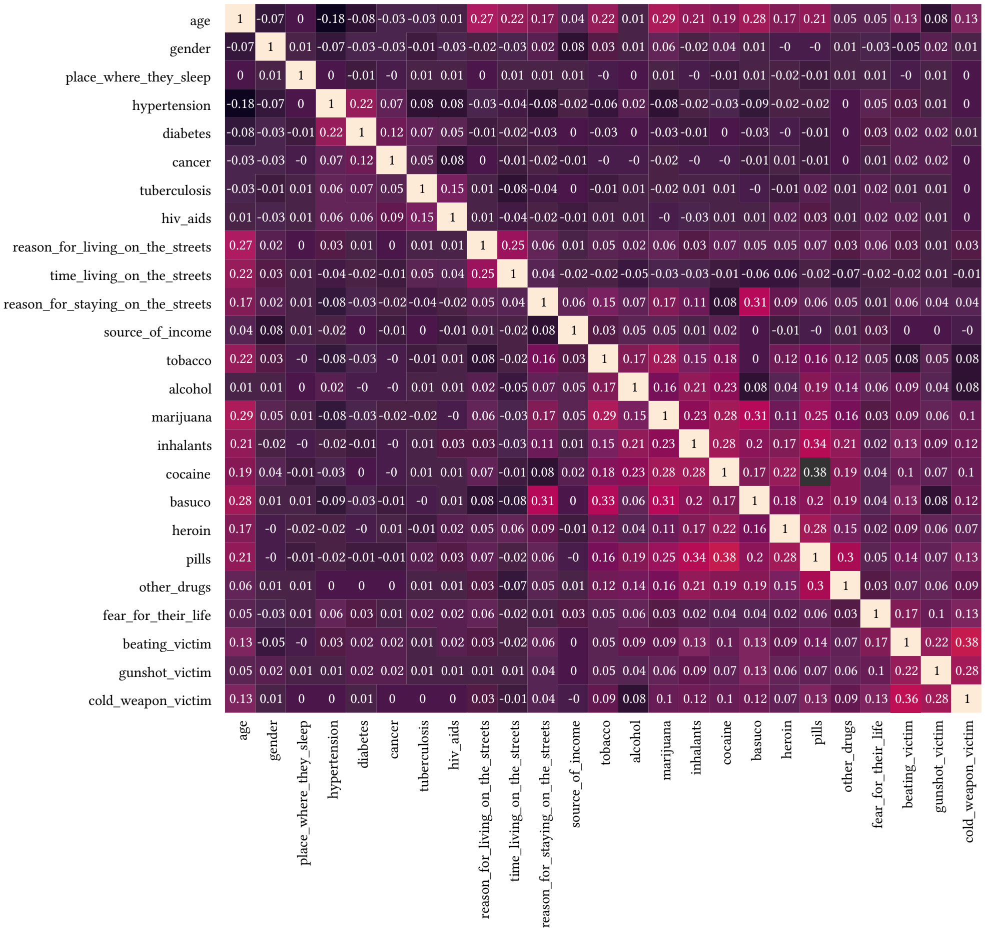


Fig. 7. Correlation (corr) among the variables of the study.

1. Data Cleaning

An exploratory analysis was carried out to understand the data. Columns or variables that did not contribute to the solution were removed, duplicate data were eliminated, and missing values were completed by data ingestion with the average between the previous and the next value in each column. Then, the records that still contained null values were eliminated. Finally, in the case of regression, the original data were normalized with the min-max method, transforming the values into a range between zero and one. The distributions of the variables and the patterns they presented were analyzed, and the relationships among the variables were identified.

2. Feature Selection

This process was carried out to achieve greater interoperability, reducing the computational cost during training and prediction, and avoid overtraining. Thus, some variables that did not contribute to the study were deleted, such as mobility, decision autonomy, personal interactions, among others. The Pearson correlation was used to quantify the linear dependence among the variables, given that the dataset includes continuous data.

Fig. 7 shows the ten variables that have the highest correlation with the study variables, showing that homeless people who use basuco, marijuana, cocaine, and heroin have a close relationship with the use of other substances, which is directly associated with their age and the reason or motive for living on the street. Regarding the test set and the training test, these were divided into training and tests, setting 70% for the first and 30% for the second.

B. Machine Learning Models

For all machine learning algorithms, hyperparameter optimization was performed with the scikit-learn library, using RandomizedSearchCV, since it is possible to obtain results as accurate as those obtained with GridSearchCV, although with a significant reduction in time, due to the sampling of the hyperparameters in the defined distribution. Cross-validation (RepeatedKFold) was also used to improve the estimated performance of each model and avoid overtraining. The data were randomly divided into subsets and optimized with the loss function: mean squared error.

- Random forest regressor: We experimented with different values for the number of trees, the number of features to consider in each split, the maximum number of levels in the tree, the minimum number of samples required to split a node, the minimum number of samples required at each leaf node, and the sample selection method to train each tree [31].
- Lasso regressor: The sum of the absolute values of the penalty weights was analyzed for the parameter n_samples, which is the number of observations to analyze the performance of the regressor [32]. Ridge regressor: The alpha was evaluated with a value of one, equivalent to an ordinary least square, and the tolerance for optimization with small values. The seed of the pseudorandom number generator was random to generate a random coefficient at each iteration. A positive value of one was used for the alpha hyperparameter to increase the variance of the estimations.

In addition, the maximum number of 15,000 iterations was defined for the conjugate gradient solver. Regarding the solver parameter, the solver was used automatically depending on the type of data [33].

TABLE II. Evaluation Metrics

Metric	Equation	Description
MSE	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$ (1)	It measures how close the points of the predictions made are to the regression line, according to the risk corresponding to the expected value of the squared error loss. The closer to zero, the better the performance of the model.
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - x_i)^2}$ (2)	It measures the difference between the actual values and those estimated by the models. The closer to zero, the better the performance of the model.
R2	$1 - \frac{\Sigma(y_i - x_i)^2}{\Sigma(x_i - \underline{x_i})^2}$ (3)	It establishes how well the actual data approximate the regression line. The closer to one, the better the performance of the model.
MAE	$\frac{1}{n}\sum_{i=1}^{n} \lvert y_i - y_i \rvert$ (4)	It calculates the errors between the actual values and those predicted by the model. The closer to zero, the better the performance of the model.
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	It measures the fraction of predictions that the model made correctly. The closer to one, the better the performance of the model.
Sensitivity	$\frac{TP}{TP + FN}$	It measures the ability of a model to detect positive cases, i.e., the predictions actually made. A high value means the model correctly identifies most of the positive results.
Specificity	$\frac{TN}{TN + FP}$	It measures the proportion of true negatives identified correctly by the model. A high value means that the model correctly identifies most of the negative results.

1. Ensemble of Machine Learning Algorithms

In this study, an ML model ensemble strategy that significantly improves the predictions made by each one separately was implemented to optimize the efficiency of the model since it combines the predictions of multiple algorithms.

The ML ensembles can be classified into three types: 1) Average ensemble, which averages the results of several models to obtain a more accurate one. 2) Voting ensemble, in which the most accurate result is obtained by counting the votes of the individual models. 3) Blending ensemble, which combines the results of the individual models to obtain a more accurate one.

In turn, the three main classes of dataset learning methods are boosting, bagging, and stacking [34]. The first two were implemented in this model.

- Boosting regressor: It is formed by a set of individual decision trees trained sequentially where each new tree tries to improve the errors of the previous trees. The prediction of a new observation is obtained by adding the predictions of all the individual trees of the model. It has the advantages of automatically selecting predictors, can be applied to regression and classification problems, handles both numerical and categorical predictors without creating dummy variables, and is not significantly influenced by outliers [35].
- Bagging regressor: It allows for better predictive performance compared to one model. The goal is to learn from a set of predictors (experts) and allow them to vote. It decreases the variance of one estimation since it combines several estimations from different models. Thus, the result can be a more stable model. Bagging is a homogeneous model of weak listeners that learn from each other independently in parallel and are combined to determine the average [36].

C. Evaluation

To evaluate the performance of the proposed model, the metrics used were mean squared error (MSE), mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R2), accuracy, sensitivity, and specificity. Table II presents the equations, the description, and the performance criterion of each evaluation metric.

TP is a result where the model correctly predicts the positive class, and TN is where the model correctly predicts the negative class. FP is a result where the model incorrectly predicts the positive class, and FN is where the model incorrectly predicts the negative class. These data were extracted from the confusion matrix.

IV. Evaluation and Results

TP is a result where the model correctly predicts the positive class, and TN is where the model correctly predicts the negative class. FP is a result where the model incorrectly predicts the positive class, and FN is where the model incorrectly.

A. Evaluation of the Regression Models

After analyzing the relationships among the variables to be predicted, the RMSE metric was used to evaluate which of the regression algorithms (random forest regression, Lasso regression, and ridge regression) presented the best results in order to select the ensemble method.

Table III presents the results of each regression algorithm. It can be seen that random forest regression provided the best results for each of the analyzed variables. The RMSE values achieved by this model are the lowest, i.e., they tend more to zero, which means that the difference between the actual and the predicted values is low. Therefore, the predictions are closely related to the actual data.

TABLE III. Regression Algorithms Evaluation RMSE Metric

Algorithm	Basuco	Marijuana	Cocaine	Heroin
Random forest regressor	**0.3898**	**0.4255**	**0.3780**	**0.2566**
Lasso regressor	0.4228	0.4582	0.4184	0.2686
Ridge regressor	0.4239	0.4584	0.4184	0.2697

Once the base algorithm for the ensemble was selected, all the regressors were evaluated. Thus, the boosting and bagging regressors were added to the evaluation. The results are presented in Table IV. It can be seen that the boosting regressor had the best results, achieving an improvement of the variables in comparison to the random forest regressor: 0.0260 basuco, 0.0154 marijuana, 0.0168 cocaine, and 0.0150 heroin.

Moreover, the results of the variables in comparison to the bagging regressor were: 0.0113 basuco, 0.01234 marijuana, 0.01073 cocaine, and 0.01325 heroin. This means that the sequential ensemble made by the boosting regressor, in which each new tree tries to improve the errors of the previous ones, allows to have optimal prediction results since, according to the metric, they tend more and more to zero, with low difference between the actual values and the predicted ones, unlike the predictions made by the random forest regressor that only considers one solution for the regression. Similarly, the boosting regressor sequential ensemble achieved better results than the bagging regressor parallel ensemble.

TABLE IV. Regressor Evaluation

Algorithm	Basuco	Marijuana	Cocaine	Heroin
Random forest regressor	0.3898	0.4255	0.3780	0.2566
Lasso regressor	0.4228	0.4582	0.4184	0.2686
Ridge regressor	0.4239	0.4584	0.4184	0.2697
Boosting regressor	**0.3638**	**0.4100**	**0.3612**	**0.2417**
Bagging regressor	0.3751	0.4224	0.3719	0.2549

B. Evaluation of the Ensemble Model

Once it was identified that the boosting regressor achieved the best results, the ensemble was evaluated with the remaining set of metrics for each of the correlated variables. Table V presents the results obtained by the proposed ensemble model. Thus, concerning the MAE metric, the best results were obtained for the heroin variable, which means that the range of error in the prediction is low since the value is very close to zero. Likewise, for the rest of the variables it could be noted that the values obtained tended to zero, demonstrating that the prediction errors of the proposed model are low.

For the MSE metric, the results obtained show that the predictions made are close to the regression line, i.e., they are closely related since all the values obtained with this metric are very close to zero.

Regarding the R2 metric, the values obtained tend to one, demonstrating that in the predictions made by the proposed model, the actual data approximate well to the regression line, with the highest point value being 0.8631 for the cocaine variable and the lowest value being 0.7674 for the basuco variable.

In the accuracy metric, the obtained values are very close to one, indicating that the fraction of predictions that the model made correctly is high, which demonstrates the excellent performance of the model. The sensitivity metric showed that the highest values were obtained for the basuco and marijuana variables, i.e., the model detected a greater proportion of positive cases for these two variables, as opposed to the cocaine and heroin variables, where the value of the metric was lower.

Finally, for the specificity metric, contrary to the previous ones, the highest values were presented in the cocaine and heroin variables since the model correctly identified a higher percentage of true negatives.

TABLE V. Results of the Evaluation of the Ensemble Model Proposed

Metric	Basuco	Marijuana	Cocaine	Heroin
MAE	0.0306	0.0247	0.0203	0.0123
MSE	0.0025	0.0013	0.0013	0.0006
R2	0.7674	0.8142	0.8631	0.7948
Accuracy	0.8193	0.7326	0.8349	0.9275
Sensitivity	0.9554	0.9194	0.3936	01966
Specificity	0.3785	0.3331	0.9664	0.99545
Matrix Confusion [[TP, FP] [FN TN]]	[[9999 467]	[[8812 772]	[[1237 1906]	[[229 936]
	[[2008 1223]]	[[2743 1370]]	[[355 10199]]	[[57 12475]]

After analyzing the performance of the model, the results of the predicted consumption of illicit psychoactive substances, according to the age of the homeless people, were plotted. Regarding the basuco variable, Fig. 8 shows that the highest use of this substance begins, on average, at 18 years old, having a high concentration between 22 and 60 years old. According to the predictions, it can also be observed that the older the person is, the lower the consumption of this substance. A possible explanation is that an older homeless person may face greater difficulties in obtaining the money to sustain this addiction.

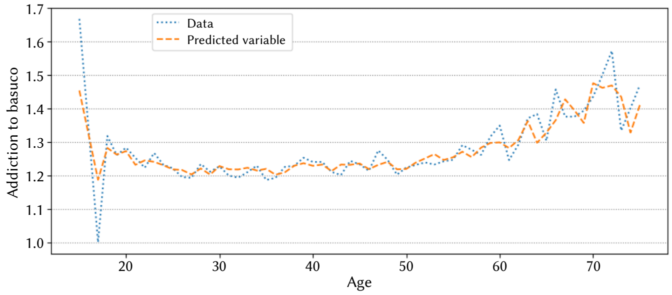


Fig. 8. Tendency to basuco addiction among homeless people according to their age.

Similarly, Fig. 9 presents the predictions made by the model regarding marijuana. It shows that most homeless people addicted to this substance are between 18 and 30 years of age, indicating that it has a greater preference among young people. Although marijuana use is not criminalized in Colombia, the number of homeless people over 40 who choose this drug is low. The direct relationship of the predictions of the proposed model with the reality of the data is remarkable.

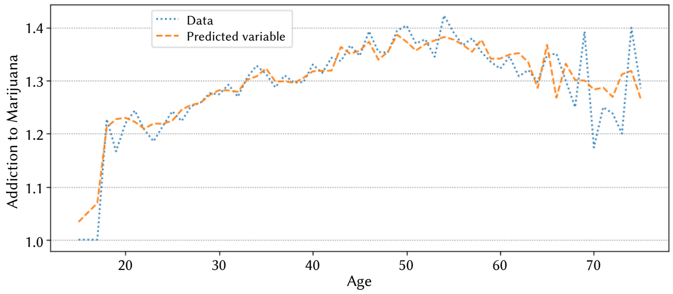


Fig. 9. Tendency to marijuana addiction among homeless people according to their age.

Fig. 10 presents the tendency to cocaine addiction of homeless people in Colombia. Most of the addicts are between 18 and 30 years of age. Its consumption decreases with the increasing age of homeless people due to its high cost, the difficulty of obtaining it on the street, and the fact that its trafficking is heavily penalized.

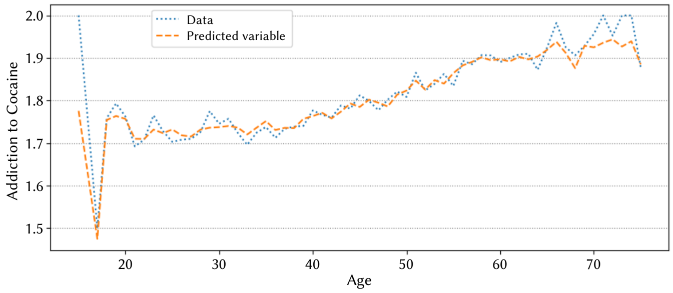


Fig. 10. Tendency to cocaine addiction among homeless people according to their age.

Regarding heroin, Figure 11 presents the prediction tendencies. This substance is similar to cocaine in that the highest concentration of addicts is found between the ages of 14 and 30 and decreases for older homeless people, for the same reasons as cocaine.

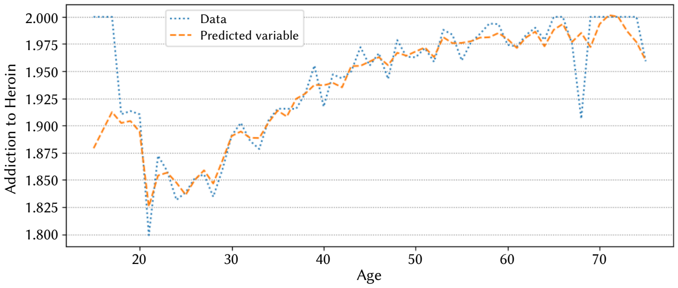


Fig. 11. Tendency to heroin addiction among homeless people according to their age.

The results of the evaluation of the applied metrics and the figures presented demonstrate that the predictions of the proposed model are close to the reality of the data.

V. Conclusions and Future Work

In this paper, an ML algorithm ensemble method was used. It was composed of three predictor algorithms and two ensemble methods. A dataset of the CHC from 2021 of the DANE containing 19,375 records and 25 columns was used to define the method. The Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology was applied to develop the research and process the data, which made it possible to understand the data set and conceptualize the data domain.

The research process allows us to conclude that not all homeless people are criminals, as many believe. Similarly, it was identified that the main sources of income of these people are recycling, cleaning car windows, and begging, which are directly related to inequality and poverty.

The study also revealed that one of the main reasons these people live on the streets is a heavy dependence on illicit psychoactive substances, with basuco being the primary substance used, followed by marijuana and cocaine. Moreover, a high number of homeless people have multiple addictions, i.e., they are addicted to more than one illicit psychoactive substance at the same time; they use regularly basuco, marijuana, and cocaine. Due to these addictions, another significant characteristic of their daily life is suffering frequent assaults, such as beatings and cold weapons injuries.

Through the evaluation of the ensemble method, it was possible to identify that boosting regressor improves performance because the individual decision trees are trained sequentially. This reduces the errors of the predecessor trees, and the prediction is made automatically based on the forecasts made by all the individual trees that make up the model.

The results show that age is closely related to addiction to certain types of illicit psychoactive substances. For example, the use of basuco is intensely concentrated in the population of homeless people between 20 and 60 years of age, that is, between youth and maturity; marijuana is consumed between 18 and 30 years of age, which indicates that younger people strongly prefer it. Although cocaine and heroin addiction occur to some extent among homeless people, it is not as prevalent because of the high cost of these substances and the difficulty of obtaining them on the street.

Finally, it can be concluded that the proposed model can serve as a decision-making tool for governmental institutions in formulating, managing and evaluating policies, plans and programs of local and municipal administrations regarding the comprehensive care, rehabilitation, and social inclusion of homeless people in Colombia.

Future work is expected to test the method with other datasets, such as access to education or home ownership in Colombia and data found in the national open data system (ANDA). In addition, the proposed method could be complemented with other ensemble techniques, e.g., stacking, which presents a general procedure for assembling base models.

References

[1] P. C. Rosa, “Exclusiones del espacio público de los habitantes de la calle en la ciudad de Buenos Aires,” *Territorios*, no. 39, p. 157, 2018, doi: 10.12804/revistas.urosario.edu.co/territorios/a.5632.

[2] P. Ruisoto and I. Contador, “The role of stress in drug addiction. An integrative review,” *Physiology and Behavior*, vol. 202, pp. 62-68, 2019, doi: 10.1016/j.physbeh.2019.01.022.

[3] G. F. Koob and J. Schulkin, “Addiction and stress: An allostatic view,” *Neuroscience and Biobehavioral Reviews*, vol. 106, pp. 245-262, 2019, doi: 10.1016/j.neubiorev.2018.09.008.

[4] E. Kpelly, S. Schauder, J. Masson, C. K. Kokou-Kpolou, and C. Moukouta, “Influence of attachment and psychotrauma in drug addiction,” *Annales Médico-psychologiques*, vol. 180, no. 6, pp. S81-S87, 2022, doi: 10.1016/j.amp.2020.11.019.

[5] I. Sadło, E. Guz, A. Wójciuk, M. Brodowicz-Król, M. Kaczoruk, and P. Kaczor-Szkodny, “Addiction to psychoactive substances as a public health challenge,” *Medycyna Ogólna i Nauki o Zdrowiu*, vol. 27, no. 1, pp. 70-76, 2021, doi: 10.26444/monz/133713.

[6] A. A. Moustafa et al., “The relationship between childhood trauma, early-life stress, and alcohol and drug use, abuse, and addiction: An integrative review,” *Current Psychology*, vol. 40, pp. 579-784, 2021, doi: 10.1007/s12144-018-9973-9.

[7] C. Rudin and K. L. Wagstaff, “Machine learning for science and society,” *Machine Learning*, vol. 95, pp. 1-9, 2014. doi: 10.1007/s10994-013-5425-9.

[8] C. Alexopoulos, V. Diamantopoulou, Z. Lachana, Y. Charalabidis, A. Androutsopoulou, and M. A. Loutsaris, “How machine learning is changing e-government,” in *ACM International Conference Proceeding Series*, 2019, pp. 354–363, 2019, doi: 10.1145/3326365.3326412.

[9] P. Cadahia, A. Golpe, J. M. Martín-Álvarez, and E. Asensio, “Measuring anomalies in cigarette sales using official data from Spanish provinces: Are the anomalies detected by the Empty Pack Surveys (EPSs) used by Transnational Tobacco Companies (TTCs) the only anomalies?,” *Tobacco Induced Diseases*, vol. 19, e98, doi: 10.18332/tid/143321.

[10] A. Andueza, M. Á. Del Arco-Osuna, B. Fornés, R. González-Crespo, and J. M. Martín-Álvarez, “Using the Statistical Machine Learning Models ARIMA and SARIMA to Measure the Impact of Covid-19 on Official Provincial Sales of Cigarettes in Spain,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 1, pp. 73-87, 2023, doi: 10.9781/ijimai.2023.02.010.

[11] A. Suruliandi, T. Idhaya, and S. P. Raja, “Drug Target Interaction Prediction Using Machine Learning Techniques – A Review,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 6, pp. 86-100, 2024, doi: 10.9781/ijimai.2022.11.002.

[12] C. F. Pereira, D. de Vargas, and L. S. Beeber, “An anxiety management intervention for people with substance use disorders (ITASUD): An intervention mapping approach based on Peplau’s theory,” *Frontiers in Public Health*, vol. 11, e1124295, 2023, doi: 10.3389/fpubh.2023.1124295.

[13] J. Ochieng, “Prevalence of Psychoactive Substance Use and Associated Behavioral Risks among Secondary School Students in Tanzania,” *East African Journal of Education and Social Sciences*, vol. 3, no. 4, pp. 185–196, 2022, doi: 10.4314/eajess.v3i4.211.

[14] J. J. Vázquez, S. Panadero, and I. Pascual, “The Particularly Vulnerable Situation of Women Living Homeless in Madrid (Spain),” *The Spanish Journal of Psychology*, pp. 1-9, 2019, doi: 10.1017/sjp.2019.58.

[15] L. C. M. Campos, J. F. de Oliveira, C. Porcino, M. J. de O. U. Reale, M. V. S. Santos, and M. E. F. de Jesus, “Social Representations Held By Homeless Individuals Regarding Homeless Individuals Who Consume Drugs,” *Revista Baiana de Enfermagem*, vol. 33, pp. 1–9, 2019, doi: 10.18471/rbe.

v33.26778.

[16] T. Coombs, T. Ginige, P. Van Calster, A. Abdelkader, O. Corazza, and S. Assi, "New Psychoactive Substances in the Homeless Population: A Cross-Sectional Study in the United Kingdom," *International Journal of Mental Health and Addiction*, e0123456789, 2023, doi: 10.1007/s11469-022-00988-7.

[17] J. J. Vázquez, A. E. Berríos, and A. C. Suarez, "Health, disability, and consumption of psychoactive substances among people in a homeless situation in León (Nicaragua)," *Social Work in Health Care*, vol. 59, no. 9-10, pp. 694-708, 2020, doi: 10.1080/00981389.2020.1835785.

[18] J. J. Vázquez, A. Suarez, A. Berríos, and S. Panadero, "Characteristics and needs of people living homeless in León (Nicaragua): Similarities and differences with other groups in severe social exclusion," *International Social Work*, vol. 65, no. 2, pp. 328–342, 2022, doi: 10.1177/0020872819896820.

[19] E. A. Salazar et al., "Inhabitants of the street in Colombia, some elements of your health," *Palarch's Journal of Archaeology of Egypt/Egyptology*, vol. 18, no. 8, pp. 3470–3476, 2021.

[20] S. Farigua, J. Pedraza, and R. Ruiz, "Experiencias de habitantes de calle que asisten al Programa de Salud Camad Rafael Uribe Uribe en Bogotá," *Revista Ciencias de la Salud*, vol. 16, no. 3, pp. 429–446, 2018.

[21] R. C. Fiorati, R. Y. D. Carretta, L. M. Kebbe, B. L. Cardoso, and J. J. D. S. Xavier, "Social ruptures and the everyday life of homeless people: an ethnographic study," Revista Gaucha de Enfermagem, vol. 37, e72861, 2017.

[22] J. C. Cubillos Álzate, M. Cárdenas, and S. Perea, *Boletines Poblacionales: Personas Adultas Mayores de 60 años Oficina de Promoción Social Ministerio de Salud y Protección Social*, Ministerio de Salud, Bogotá D. C., 2020. Available: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/PS/boletines-poblacionales-envejecimiento.pdf [Accessed June 10, 2023]

[23] H. D. Whitehead et al., "Validated method for the analysis of 22 illicit drugs and their metabolites via liquid chromatography tandem mass spectrometry (LC-MS/MS) in illicit drug samples collected in Chicago, IL," *Forensic Chemistry*, vol. 33, e100475, 2023, doi: 10.1016/j.forc.2023.100475.

[24] D. Yajaira, B. Fernández, Á. Segura-Cardona, L. Montoya-Velez, and M. Hernández-Rendón, "Consumo de basuco en usuarios de drogas inyectables en Colombia," *Revista Cubana de Salud Pública*, vol. 42, no. 2, pp. 276-283, 2016.

[25] A. H. Sadaka et al., "Effects of inhaled cannabis high in Δ9-THC or CBD on the aging brain: A translational MRI and behavioral study," *Frontiers in Aging Neuroscience*, vol. 15, pp. 1-20, 2023, doi: 10.3389/fnagi.2023.1055433.

[26] S. E. Koch, J. A. Marckel, J. Rubinstein, and A. B. Norman, "A humanized anti-cocaine mAb antagonizes the cardiovascular effects of cocaine in rats," *Pharmacology Research and Perspectives*, vol. 11, no. 1, pp. 1–8, 2023, doi: 10.1002/prp2.1045.

[27] F. Gosetti, et al, "From the Streets to the Judicial Evidence: Determination of Traditional Illicit Substances in Drug Seizures by a Rapid and Sensitive UHPLC-MS/MS-Based Platform," *Molecules*, vol. 28, no. 1, e164, 2022, doi: 10.3390/molecules28010164.

[28] P. J. Cooper et al., "Understanding and controlling asthma in Latin America: A review of recent research informed by the SCAALA programme," *Clinical and Translational Allergy*, vol. 13, no. 3, e12232, 2023, doi: 10.1002/clt2.12232.

[29] Ministerio de Salud y Protección Social, *Política Integral para la Prevención y Atención del Consumo de Sustancias Psicoactivas*, p. 44, 2019. Available: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/PP/politica-prevencion-atencion-spa.pdf [Accessed June 11, 2023]

[30] G. Gesti, *Política Pública Social para Habitantes de la Calle 2021-2031*, p. 231, 2021. https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/PS/abece-habitantes-calle-2022-2031.pdf [Accessed June 11, 2023]

[31] S. Kwak et al., "Machine learning prediction of the mechanical properties of γ-TiAl alloys produced using random forest regression model," *Journal of Materials Research and Technology*, vol. 18, pp. 520-530, 2022, doi: 10.1016/j.jmrt.2022.02.108.

[32] P. Maranzano, P. Otto, and A. Fassò, "Adaptive LASSO estimation for functional hidden dynamic geostatistical model," *Stochastic Environmental Research and Risk Assessment*, vol. 37, pp. 3615-3637, 2022, doi: 10.1007/s00477-023-02466-5.

[33] S. Mohammadi, "A test of harmful multicollinearity: A generalized ridge regression approach," *Communications in Statistics - Theory and Methods*, vol. 51, no. 3, pp. 724-743, 2022, doi: 10.1080/03610926.2020.1754855.

[34] B. Das et al., "Comparison of bagging, boosting and stacking algorithms for surface soil moisture mapping using optical-thermal-microwave remote sensing synergies," *Catena*, vol. 217, e106485, 2022, doi: 10.1016/j.catena.2022.106485.

[35] M. Sipper and J. H. Moore, "AddGBoost: A gradient boosting-style algorithm based on strong learners," *Machine Learning with Applications*, vol. 7, e100243, 2022, doi: 10.1016/j.mlwa.2021.100243.

[36] P. W. Khan, S. J. Park, S. J. Lee, and Y. C. Byun, "Electric Kickboard Demand Prediction in Spatiotemporal Dimension Using Clustering-Aided Bagging Regressor," *Electric Vehicles: Planning and Operations*, vol. 2022, e8062932, 2022, doi: 10.1155/2022/8062932

Hugo Ordoñez

Full-time professor at the University of Cauca. PhD in Telematics Engineering Universidad del Cauca 2015, Master in Computing Universidad del Cauca 2011. Area of Interest: Business process discovery, information management.

Silvio Ricardo Timarán-Pereira

Full-time professor at the University of Nariño. Systems Engineer and Master Of Science in Engineering from Donetsk Polytechnic University.

Juan-Sebastián González-Sanabria

Systems and Computing Engineer, from UPTC Tunja, and has two specializations: one in Databases at UPTC and another from the National University of La Plata in Scientific and Technological Information Management. In addition, he has a Master's Degree from the International University of La Rioja in Software and Information Systems. In his teaching activity, he has worked in the subjects of Databases and Research. He has more than a dozen papers and scientific articles in different journals, mainly in the area of data analysis and development.

Use of Optimised LSTM Neural Networks Pre-Trained With Synthetic Data to Estimate PV Generation

Miguel Martínez-Comesaña[1]*, Javier Martínez-Torres[2,3], Pablo Eguía-Oller[1], Javier López-Gómez[1]

[1] Department of Mechanical Engineering, Heat Engines and Fluids Mechanics, Industrial Engineering School, CINTECX, University of Vigo (Universidade de Vigo), Maxwell s/n, 36310 Vigo (Spain)
[2] Department of Applied Mathematics I, Telecommunications Engineering School, CINTECX, University of Vigo (Universidade de Vigo), 36310 Vigo (Spain)
[3] Department of Applied Mathematics I, Telecommunications Engineering School, CITMAga, 15782 Santiago de Compostela (Spain)

* Corresponding author: migmartinez@uvigo.gal

Received 13 September 2022 | Accepted 17 October 2023 | Published 17 November 2023

Abstract

Optimising the use of the photovoltaic (PV) energy is essential to reduce fossil fuel emissions by increasing the use of solar power generation. In recent years, research has focused on physical simulations or artifical intelligence models attempting to increase the accuracy of PV generation predictions. The use of simulated data as pre-training for deep learning models has increased in different fields. The reasons are the higher efficiency in the subsequent training with real data and the possibility of not having real data available. This work presents a methodology, based on an deep learning model optimised with specific techniques and pre-trained with synthetic data, to estimate the generation of a PV system. A case study of a photovoltaic installation with 296 PV panels located in northwest Spain is presented. The results show that the model with proper pre-training trains six to seven times faster than a model without pre-training and three to four times faster than a model pre-trained with non-accurate simulated data. In terms of accuracy and considering a homogeneous training process, all models obtained average relative errors around 12%, except the model with incorrect pre-training which performs worse.

Keywords

Genetic Algorithm, LSTM, Optimisation, Pre-Training, PV Power, Synthetic Data.

DOI: 10.9781/ijimai.2023.11.002

I. Introduction

Nowadays, the demand for electric power is growing significantly and the mayor issue is to reduce fossil fuel emissions and thus control global warning [1]. Transport and electricity generation have accounted for 60% of all energy produced in the last few years [2]. In this way, the European Commission has defined new targets for 2030 which include reducing the CO_2 emissions by 40% with respect to 1990 levels [3]. Meeting this target requires reducing the electricity demands and/or increasing the use of renewable energies [4].

Among the renewable energies, solar power generation has proven to be a serious option as a result of its great availability and low production cost [5]. This type of renewable energy generation has two main sources: thermal and photovoltaic (PV). In recent years, solar PV production has expanded considerably, becoming the fastest growing resource for electric power generation with the highest power density among all renewable energy resources [6]–[8]. This resource also has two important barriers: the low efficiency of the PV modules (directly related to meteorological conditions) and the large investment cost [5], [9]. Nevertheless, its potential to feed energy into the grid along with the reduction of transmission losses it provides makes this renewable resource very attractive [10].

Recently, artificial intelligence techniques, more specifically deep learning models, has become widespread as a novel data-driven approach that can be applied to numerous scientific fields such as PV energy analysis or related areas [11], [12]. Deep learning models are famous because they are able to learn complex patterns without requiring in-depth knowledge of the subject under analysis and are characterised for their high performance and easy implementation. In addition, these models have become increasingly more popular due to their ability to better optimise and replicate learning patterns than the more classical machine learning techniques [11]. Some concrete examples of that, in similar studies of the one proposed, are Nabipour et al. [13] show the higher accuracy of DL models prediction stock market trends and Mert. [14] show the better performance of DL models in solar-powered systems production estimations.

Please cite this article in press as:
M.Martínez-Comesaña, J. Martínez-Torres, P. Eguía-Oller, J. López-Gómez. Use of Optimised LSTM Neural Networks Pre-Trained With Synthetic Data to Estimate PV Generation, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 61-70, 2025, http://dx.doi.org/10.9781/ijimai.2023.11.002

Long Short-Term Memory (LSTM) neural networks are a deep learning model, within the group of Recurrent Neural Networks (RNN) [15], which contain a specific hidden layer that considers the existence of connections with past values [16], [17]. In this way, they are suitable for mapping long-term dependencies. These neural networks have been implemented in similar fields such as environment [18], energy efficiency in buildings [16], image processing [17] and PV generation [19], [20]. In particular, they have shown better performance in photovoltaic generation estimations thanks to being able to use the information learned from previous steps [21], [22]. Moreover, most deep learning models leave room for optimisation based on the hyperparameters that defined them. These improvements, which can be achieved in model performance reaching the optimal values for the hyperparameters is demonstrated in several previous studies [23],[24]. In the literature can be found different techniques to efficiently perform this search: univariate dynamic encoding algorithms [25], combination between grid and random searches [26], particle swarm optimisation [27] or genetic algorithms [28]. In particular, Genetic Algorithms (GA) have increased their use in this type of optimisations mainly due to their easy of implementation and the reduction in the number of evaluations and time needed to reach an optimum [24], [29], [30]. Furthermore, multiobjective genetic algorithms such as Non-dominant Sorting Genetic Algorithm (NSGA-II) make it possible to optimise the values of the selected hyperparameters considering more than one objective function [1], [31], [32].

In recent years, feeding machine and deep learning models with simulated data, prior to real data, has been shown to improve their performance. The spread of this technique is due to the fact that several studies have shown that pre-training the model with synthetic data enables subsequent training with real data to be faster and/or more accurate, on the one hand, and that in certain situations collecting real data is very costly or not possible, on the other hand [33], [34]. This methodology has been applied in several fields such as signal denoising [35], pattern recognition [36] and robot perception [37]. The aim of this research is to introduce a methodology to optimise deep learning models architecture and improve their performance using synthetic data. In particular, this study focuses on estimating PV generation and comparing the accuracy of the built models depending on their pre-training. The analysed installation is located on the roof of the School of Mining Engineering in northwest Spain. The available data consist of hourly frequency observations of PV generation together with outdoor temperature and global solar irradiance of the area. Additionally, three temporal variables (month of the year, day of the month and hour of the day) are also considered as model inputs. In this way, taking into account the aforementioned inputs and the variable of interest (PV generation), the optimisation process and the improvement provided by a proper pre-training were analysed. Specifically, both the epochs required to reach a certain error limit and the coefficient of variation of the root mean squared error (CV(RMSE)) and normalised mean bias error (NMBE) are the model evaluation metrics selected.

The novelty of this paper lies in the application of deep learning models, optimised with the NSGA-II algorithm and pre-trained with simulated data, to perform PV generation predictions of an installation consisting of 296 PV modules. Furthermore, the introduced methodology shows the significant improvement of the model behaviour with a correct pre-training process based on synthetic data. Thus, this work contributes with a method that efficiently optimises the deep learning model and improves its training speed in comparison with a model without pre-training or with an incorrect pre-training. In the field of renewable energies, this improvement allows better control and utilisation of photovoltaic energy, optimising, for example, the connection between a house with photovoltaic panels and an electric vehicle.In addition, the presented use of synthetic data allows the implementation of deep learning models in situations where the monitored data is limited or the PV systems have just been installed and there is very few data available to feed the model.

II. Material and Methods

The aim of this research is to analyse the usefulness of synthetic data to pre-train deep learning models and thus study whether they improve their performance in the training process with real data. In this case, the study focuses on a photovoltaic installation, and specifically, on estimating the generation of a PV system based on meteorological and temporal variables. To this end, the deep learning models used are LSTM neural networks optimised with NSGA-II multiobjective genetic algorithm.

A. Long Short-Term Memory (LSTM) Neural Network

The deep learning model used in this study is a Long Short-Term Memory (LSTM) neural network. This type of neural networks are Recurrent Neural Networks (RNN); sequenced-based models that take into account the possible correlations between past and current data [38], [39]. RNN use the backpropagation through time (BPTT) method, which considers that the decision a RNN makes at time step $t-1$ can influence the decision at time step *t*. However, due to the vanishing gradient problem [40], these models are not good learning relationships in the long run. This problem is described as the gradient norm decays exponentially to zero from long-range dependencies. In this case, LSTM neural networks, having an architecture with a memory cell and a forget gate, are capable of solving the aforementioned problem [41].

The dynamics of RNN can be established with deterministic transitions from previous to current hidden state (h_t^l):

$$RNN\colon \mathbf{h}_t^{l-1}, \mathbf{h}_{t-1}^{l} \rightarrow \mathbf{h}_t^l \tag{1}$$

being l the layer and t the time step. In constrast, LSTM neural networks present a more sofisticated structure that enables the memorisation of information for many time steps. The long-term memory is stored in a dedicated vector of memory cells $s_t^l \in \mathbb{R}^k$:

$$LSTM\colon \mathbf{h}_t^{l-1}, \mathbf{h}_{t-1}^{l}, \mathbf{s}_{t-1}^{l} \rightarrow \mathbf{h}_t^l, \mathbf{s}_t^l \tag{2}$$

As an illustration, we assume an input vector x, where $x_t \in \mathbb{R}^k$ is a *k*-dimensional vector at time step *t*. LSTM neural networks maintain an internal memory cell state during the entire process in order to build the temporal connections. The memory cell s_{t-1} interacts with the hidden state h_{t-1} and the specific input x_t to establish the elements of the inner state vector to be deleted, updated or mantained. Furthermore, LSTM neural networks have a forget gate f_t, an input gate i_t, an input node n_t and an output gate ot in their structure (see Fig. 1). The architecture of these models can be defined by the equations 3, 4 and 5 [38], [41]:

$$\mathbf{f}_t = \sigma(W_{fx}\mathbf{x}_t + W_{fh}\mathbf{h}_{t-1} + \mathbf{b}_f) \tag{3}$$

$$\mathbf{i}_t = \sigma(W_{ix}\mathbf{x}_t + W_{ih}\mathbf{h}_{t-1} + \mathbf{b}_i) \tag{4}$$

$$\mathbf{o}_t = \sigma(W_{ox}\mathbf{x}_t + W_{oh}\mathbf{h}_{t-1}\mathbf{b}_o) \tag{5}$$

being W weight matrices associated to the activation functions, $\odot$ an element-wise multiplication and σ the representations of the sigmoid function.

In this way, the new current cell state (n_t) can be calculated with Equation 6:

$$\boldsymbol{n}_t = \phi(W_{nx}\boldsymbol{x}_t + W_{nh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_n) \tag{6}$$

where φ represent the tanh activation function. Based on the forget and input gate the state st is updated through Equation 7:

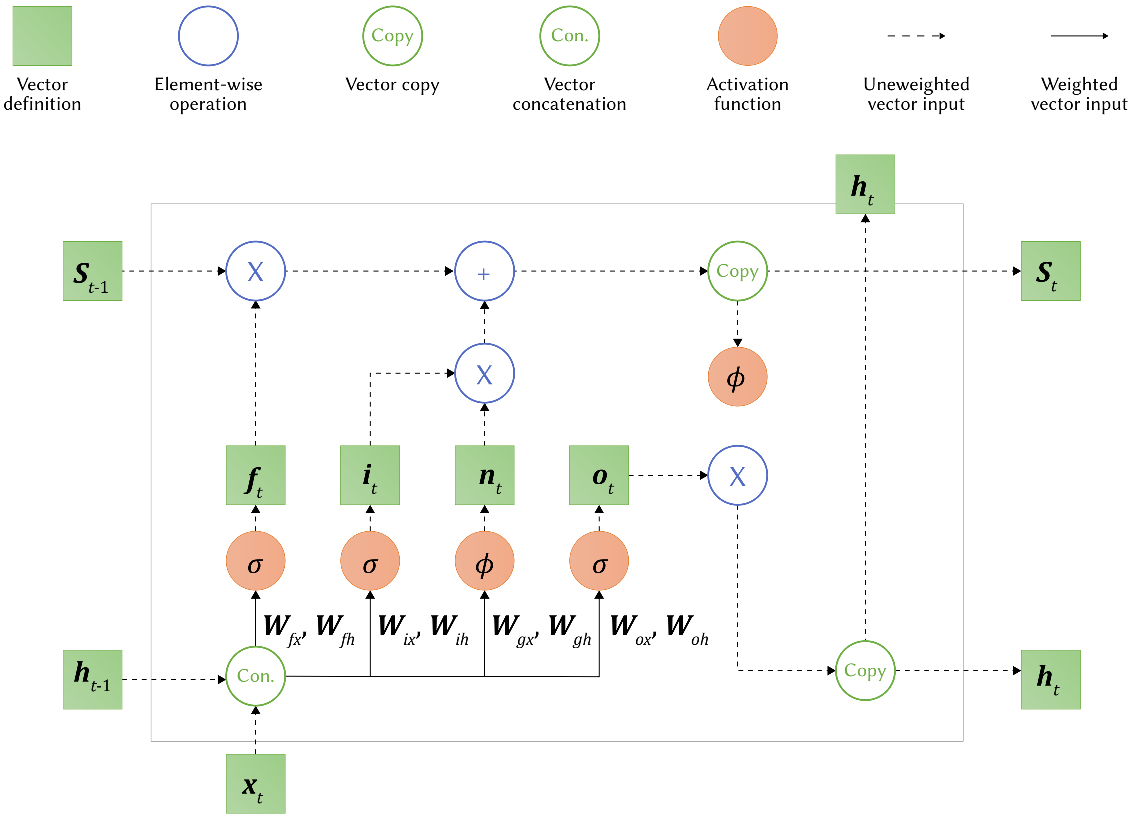


Fig. 1. Internal structure of LSTM block.

$$\boldsymbol{s}_t = \boldsymbol{n}_t \odot \boldsymbol{i}_t + \boldsymbol{s}_{t-1} \odot \boldsymbol{f}_t \tag{7}$$

and the current hidden output using Equation 8:

$$\mathbf{h}_t = \phi(\mathbf{s}_t) \odot \mathbf{o}_t \tag{8}$$

As shown in Fig. 1 there are three sigmoid functions in the LSTM block, which can be 0 or 1 and act as switches to manage which elements pass through the gates. In addition, the present input xt and the past state ht1 affect the decision made at the forget gate f, the input gate i and the output gate o. The forget gate determines which elements of the previous memory cell st1 are forgotten and the input gate selects which elements are kept. Thus, the inner state is updated and the elements of st that move forward as LSTM state ht are selected through the output gate. This process is replicated at every time step [39], [41].

On the one hand, the LSTM neural networks built in this analysis are optimised with a mutiobjective genetic algorithm focusing on the accuracy and the complexity of the model. The parameters adjusted are the number of LSTM layers, the number of Dense layers, the number of neurons in each of them and the number of epochs the model is allowed not to improve (stopping criterion). On the other hand, the built neural networks use the internal optimisation algorithm known as *Adam*, the Rectified Linear Unit (reLU) activation function and a batch size of 24.

B. Model Optimisation

The optimal architecture together with the optimal value for the parameter defining the stopping criterion (number of epochs without improvement) of the built LSTM neural network are obtained with a multiobjective genetic algorithm. Genetic Algorithms (GA) are known for trying to replicate biological evolution to solve optimisation problems. They initiate the process with a random population based on individuals. These individuals are represented by chromosomes consisting of genes which, in turn, are the values of the considered covariates. Thus, this type of algorithms conducts optimisation based on three main operators: crossover, mutation and elitism. Crossover refers to exchanging a portion of a specific chromosome with a portion of another random chromosome. Mutation increases diversity in populations to avoid stagnating at local optima by randomly modifying part of solutions. Elitism is the way in which the selection process is accomplished by choosing the best chromosomes to pass through generations [42], [43].

In this study, the specific algorithm used is the Non-Dominant Sorting Genetic Algorithm (NSGA-II). It is a robust multiobjective algorithm widely implemented in different practical fields that allows the simultaneous optimisation of several parameters. Furthermore, it is characterised by generating a Pareto front betweeen the objectives where the overall optimum is selected and for being an improved version of the original version of the NSGA. These improvements are based on the use of a crowding distance operator, the elitism and a fast nondominated ranking [31], [44].

This algorithm is based on four internal principles that defined its processing [45]:

- Non-dominated sorting: The options considered, which form a population, are ordered by Pareto dominance. In this way, the elements/options with the best rank are separated and the ordering continues with the rest of the options.
- Crowding distance: Between two possible solutions, the one with a larger crowding distance is considered to be in a less crowded area. Thus, the elements in a less crowded region will be selected first. The crowding distance for an element is presented in the Equation 9:

$$\mathrm{CD}(i) = \sum_{i=1}^{k} \frac{F_j^{i+1} - F_j^{i-1}}{F_j^{\max} - F_j^{\min}} \tag{9}$$

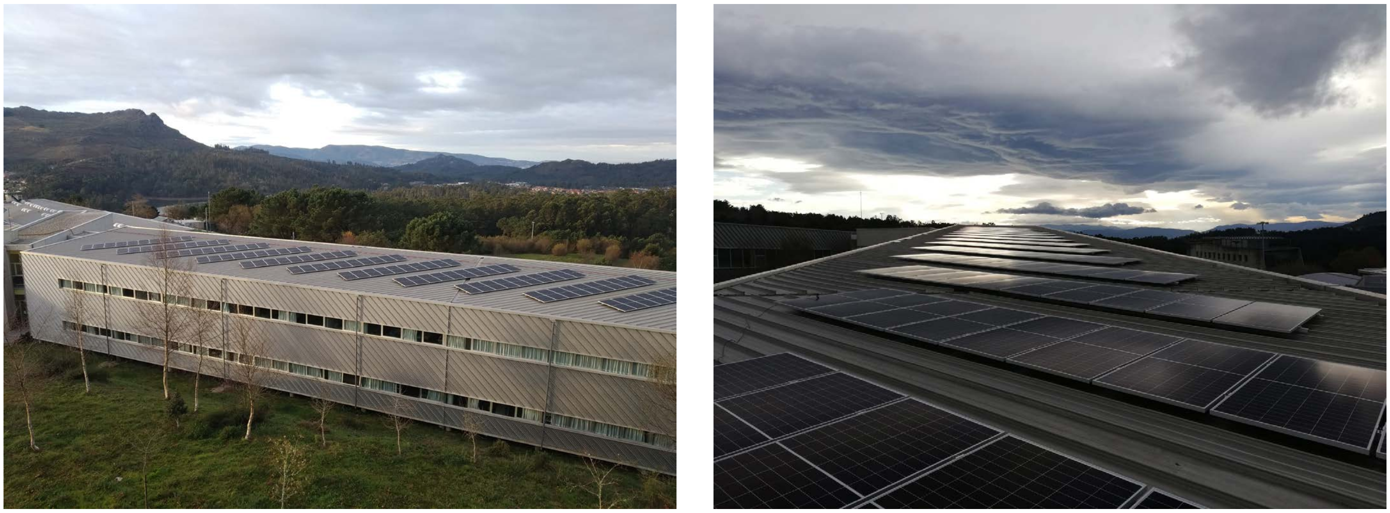


Fig. 2. Pictures of the PV installation analysed.

where k is the number of objectives, F_j^i the value of the i-th element for objective j, and $F_j^{\max}$; $F_j^{\min}$ the maximum and minimum values for objective j.

- Elitism: The best option combinations pass directly pass to next generations of the algorithm. Non-dominated combinations continue until another solution dominate them.
- Selection operator: The selection of elements to be transferred for next generations is based on their rank and their crowding distances.

The objective functions considered to be minimised with NSGA-II are the CV(RMSE) in PV generation predictions and a complexity function that summarises the layers and neurons of the model. This complexity function, already use in [31], [46], relies on the number of layers and neurons in the built neural network:

$$\text{Complexity} = 0.25 \times \frac{l}{L} + 0.75 \times \frac{\sum_{j=1}^{L} n_j}{N} \tag{10}$$

with l and L being the number of layers used and the maximum value allowed (in this analysis, 5). In addition, n_j and N represent the neurons in each layer and the maximum number of neurons allowed (in this analysis, 500). In order to avoid rejecting excessive multilayer architectures, a lower weighting for the number of layers is introduced. In this case, the termination of the optimisation process is based on a specific tolerance value within the space of feasible solutions and the optimal point along the final Pareto front is selected using a decomposition function, known as penalty boundary intersection (PBI) [47].

Further information about the NSGA-II can be found in [48].

C. Validation and Error Assessment

The validation metrics considered in this analysis to evaluate the accuracy of the deep learning models are the the Coefficient of Variation of the Root Mean Square Error (CV(RMSE)), Normalised Mean Biased Error (NMBE) and Mean Absolute Error (MAE):

$$\text{CV(RMSE)} = 100 \times \frac{\sqrt{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2/N}}{\overline{y}} \tag{11}$$

$$\text{NMBE} = 100 \times \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)}{\sum_{i=1}^{N}(y_i)} \tag{12}$$

$$\text{MAE} = \frac{\sum_{i=1}^{N}|\,y_i - \hat{y}_i\,|}{N} \tag{13}$$

where y_i represents the real values, $\hat{y}_i$ the estimations and N the number of observations. These metrics are used to compare the performance of the built LSTM neural networks throught a cross-validation process (considering an expanding window) with average results presented in the section IV. They were used in similar studies such as [24], [49], [50]. Moreover, the accuracy of the models is assessed only considering the hours with positive solar irradiance (without irradiance it is known that the panels do not produce).

III. Experimental System

The studied PV system is an installation located on the roof of the School of Mining Engineering in north-western Spain at University of Vigo (see Fig. 2).

This installation is composed by 296 PV modules in parallel, with an azimuth of 72.8°-112.6°, because two groups of modules are considered, and a slope of 2°. In addition, the specific coordinates of the installations are latitude of N 42° 10' 6.1'' and longitude of W 8° 41' 18.44''. The technical information about the inverters and PV modules of the analysed installation is presented in Table I.

TABLE I. PV Inverters and Modules Datasheets

Inverter	$V_{DC,max}$	1000 V
	$V_{DC,MPP}$	500 - 800 V
	$I_{DC,max}$	120 A
	$I_{SC,max}$	30 A
	$V_{AC,nom}$	230 V
	f_{nom}	50 Hz
	$I_{AC,max}$	72.5 A
PV module	P_{MPP}	400 W
	Clasification range	0/+5 W
	Accuracy (P_{MPP})	± 3%
	U_{MPP}	40.32 V
	I_{MPP}	9.92 A
	U_{OC}	400 V
	I_{SC}	10.45 A

A. Synthetic Data

In this study, the available simulated data is generated with the software TRNSYS [51], [52]. The data consist of simulated photovoltaic generation based on physical laws and weather data significantly correlated with PV generation (in this case outdoor temperature and solar irradiance) along one year (see Fig. 3). The aforementioned simulation, considering the same weather conditions, is carried out for different PV installations considering different number of PV modules in parallel, different azimuths and different slopes. The number of PV modules varies between 60 and 740 (60, 89, 178, 296, 414, 562, 740), the azimuth between 0 and 337.5 degrees (22.5 by 22.5) and the slope between 0 and 90 degrees (15 by 15) generating data from 784 different PV instalations. Among this grid of parameters combinations there is the same configuration as the analysed installation.

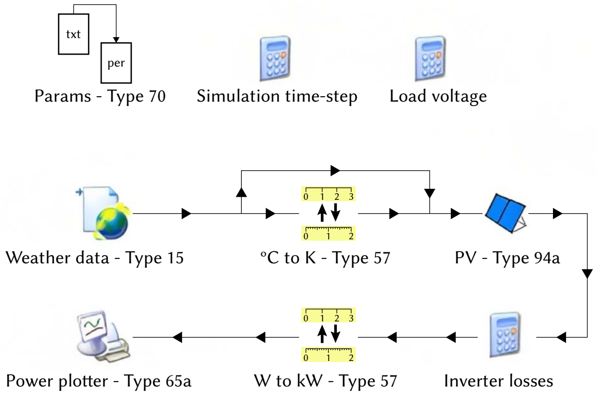


Fig. 3. Simulation process followed by TRNSYS in order to generate PV generation data.

The purpose of these synthetic data is to provide data from different installations (to fit a wide range of possibilities) in order to subsequently pre-train deep learning models and improve their performance on real data. Thus, the deep learning model reaches the training process, with real data, knowing the relationship between the selected inputs and the specific power generation of the installed panels.

B. Weather Data

The meteorological variables considered as model inputs in this analysis are global solar irradiance and outdoor temperature. They have a significant correlation with the generation of the photovoltaic modules [53]. Specifically, the data source used to obtain these data is an automatic weather station belonging to a meteorological agency known as MeteoGalicia [54]. The station is located 250 m northeast of the centre of the PV installation and 35 m higher. For missing or invalid values collected by the station, the Global Forecast System (GFS flux) surface flux model is used. This model generates hourly forecasts on a 13 km resolution grid [55].

C. Data Preprocessing

This research is focused on analising the improvement, on PV generation estimations, that produces pre-training a deep learning model with simulated data (see Fig. 4). In addition to the right installation parameters, the deep learning model is also pre-trained with simulated data based on random parameters (extracted from the list of section A) to consider the case where these data are not available. As mentioned, the aim of the built LSTM neural network is to predict the generation of a PV installation. The data available in this analysis are hourly observations of the PV generation of the studied installation and simulated observations, considering the parameters of that installation and 783 variations of them (Section A), in addition to the solar irradiance and outdoor temperature of the area. The availability of the real data corresponds to the year 2021 (from March to September) and the simulated data corresponds to 2020. In this period of time there is no missing or invalid data. Three complementary variables related with the time (hour of the day, day of the month and month of the year) are also considered as model inputs to improve the accuracy of the model. In order to take into account the existing inertia in the solar irradiance, and thus in PV generation, 24 hourly lags are considered. Moreover, the data set is normalised based on the limits 0 and 1.

As mentioned, the pretaining is conducted with simulated data considering on the one hand the parameters of the studied installations (n_panels: 296, azimuth: 90° and slope: 2°) and, on the other hand, a random set of parameters n_panels: 562, azimuth: 247.5° and slope: 60°). Specifically, these parameters are the number of modules, their azimuth and their slope. The following section presents two analyses: one focused on introducing the process of selecting the optimal

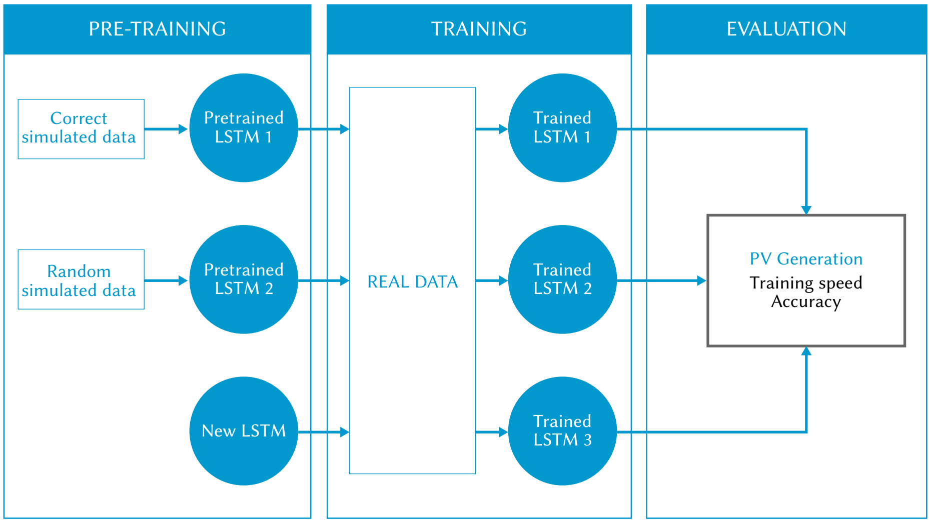


Fig. 4. Research methodology in which the three parts (pre-training with synthetic data, training with real data and the subsequent evaluation) are presented in different ways depending on the pre-training. In addition, the measures selected to compare the performance of the models, which are training speed and accuracy, are shown.

architecture and stopping criterion of the deep learning model and the other focused on showing the improvement in training speed and model accuracy due to pre-training (see Fig. 4).

IV. Results and Discussion

This paper presents a methodology for estimating PV generation optimised with a genetic algorithm that searches for the best LSTM neural network architecture together with the best stopping criterion for training and improved with pre-training based on simulated data. The inputs to produce the PV generation estimations of the built models are solar irradiance, outdoor temperature and three temporal variables. To validate this methodology, monitored data from a photovoltaic installation located in the northwest of Spain are available. On the one hand, section A shows the results and parameters used in the optimisation process of the LSTM neural network architecture and its stopping criterion. On the other hand, section B presents the improvements obtained by pre-training the model with synthetic data based on speed and accuracy. In this section, a comparison between two different pre-training and no pre-training is presented by analysing the number of epochs needed to reach certain error levels (directly related with time) and the accuracy they yield with similar training (same stopping criterion). In addition, the proposed optimisation and method, along with the following results, were implemented using the Python programming language [56].

A. LSTM Neural Network Optimisation

The optimal selection of the LSTM neural network architecture (considering LSTM hidden layers, dense hidden layers and the neurons within them) and the number of epochs without enhancement to stop training is obtained with NSGA-II. The average CV(RMSE), from a cross-validation experiment on the simulated sample corresponding to the studied system, and the complexity of the model (Equation 10) are the objective functions considered. The aim of the multi-objective genetic algorithm is to minimise these functions simultaneously. The optimisation is conducted with simulated data instead of real data, in order to assume the situation where real data is not yet available. The table II shows the specific hyperparameters used in the optimisation process: those that define the option space and those that configure the algorithm termination and selection process.

TABLE II. Parameters and Functions Used in the Optimisation Through NSGA-II, Comprising the General Parameters Related to the Multiobjective Algorithm and the Specific Parameters Related to the Optimal Selection

General parameter	Value	Termination parameter	Value
Neurons options	20 100 (20 by 20)	**Tolerance (*tol*)**	0.1
LSTM layer options	1 3	**Nº max evals (*n_max*)**	5000
Dense layer options	0 2	**Last genes considered (*n_last*)**	40
Patience options	10 or 20 epochs	**Decomposition function**	PBI
Population	50		
Mutation	0.9		
Crossover	0.1		

In this case, considering the parameters presented in Table II, NSGA-II needed 2490 evaluations to find 5 optimal points on the Pareto front (7688 possible options in total). These points correspond to LSTM architectures and epoch limits to stop the model training. Then, the PBI decomposition function taking into account heterogeneous weights (0.75{0.25), respectively for the error and complexity objective functions, is used to select a point on the Pareto front. Although in this case we give more importance to error, the distribution of weights can be adapted to obtain less accurate but simpler models.

The results are an LSTM neural network architecture with an LSTM hidden layer with 80 neurons and a Dense hidden layer with 40 neurons (5 80 40 1), as well as a model patience, measured in epochs, of 20. More information and details of this selection process can be found in [31], [57], [58].

B. LSTM Neural Network Performance

Once the optimal LSTM architecture and the stopping criterion for training the model have been obtained, two different analyses are performed: one based on analysing the improvement in training speed produced by a model pre-training and the other focused on comparing the differences in accuracy between the built models considering the same training process. In this case, the comparison is carried out considering three different models: one without pre-training, one with a random pre-training and one pre-trained with the parameters of the studied installation (number of PV modules, azimuth and slope). In this specific analysis, the values of the random parameters selected are 562 PV modules with an azimuth of 247.5° and a slope of 60°.

On the one hand, Table III, in which the training speed is analysed, shows the average results (30 repetitions) of measuring the number of epochs each model requires to reach certain error limits, also taking into account the time, measured in seconds, required to reach it. Normalised data and the Mean Squared Error (MSE), for error limits, are considered. The pre-trained models use one year of simulated data and all built models are retrained and evaluated with seven months of real data (first 4 for training and the remaining for validation).

TABLE III. Average Results of 30 Repetitions of an Experiment in Which the Number of Epochs Needed by Each Model to Reach the Error Limits Shown Are Analysed. The Average Number of Epochs and Time Each Model Needed to Reach the Limits Are Presented

MSE Limits	Correct pre-training		Random pre-training		No pre-training	
	Epochs [n]	Time [s]	Epochs [n]	Time [s]	Epochs [n]	Time [s]
0.005	1	4.86	1	4.88	3.90	8.00
0.004	1	4.41	1	4.61	5.27	9.51
0.003	1	4.45	1	4.42	8.43	12.54
0.002	3.33	7.41	18	25.51	42.67	46.11

In the case of the first limits (0.005, 0.004, 0.003), both pre-trained models with synthetic data only needed one epoch to reach the limit. The model without pre-trainig is the slowest, needing more than 3, 5 and 8 epochs on average to reach respectively the first mentioned limits. Considering the times spent on training, the pre-trained models reduce it to half at the first limit and to one third at the third limit. With regard to the last error limit, the differences between the three built models become more significant. The model with the correct pre-training requires on average 3.33 epochs to reach the error limit, while the model with a random pre-training requires 18 epochs. Moreover, the model without pre-training remains the slowest, taking, on average, 42.67 epochs. Observing the times the results are similar: the model with a correct pre-training spent, on average, 7.41 seconds (more than three times less than the model pre-train with random parameters (25.51) and more than six times less than the model without pre-training (46.11). In this way, it can be seen that the improvements, in terms of speed, provided by a pre-training with synthetic data are significant considering both correct and incorrect parameters. The information extracted in this pre-training generates models able to adapt faster to real situations, although considering the right pre-training is more efficient.

On the other hand, the results of the study of the accuracy of the built models following a homogeneous training process are presented in Table IV. The training process is based on a cross-validation experiment considering an expanding window; the models are evaluated on the seven months of real data available (one by one) using the remaining previous months for training. In addition, the LSTM architecture and the stopping criterion considered are those obtained in the previous section.

In terms of CV(RMSE), with which the average distance to the real curve is measured, it can be observed in Table IV that the average value of the model with a random pre-training is significantly higher than the others (21.15 %). The standard deviation among all CV(RMSE) values is also the highest (± 5.87), showing a large variability in the results. The average CV(RMSE) yielded by the model with correct pre-training and the one without pre-training are close, with similar variability, but the former is lower (12.84 % and 14.22 % respectively). As for the NMBE results, which measure how close the estimations are on average to reality, the model with incorrect pre-training has the highest average value (0.10 %) and the highest variability in results (0.08) (see Table 4). In this case, the model with no pre-training presents the lowest value (0.07 %) followed by the model pre-trained with correct parameters (0.07 %), both with controlled variances. Regarding the MAE results, a metric that measures the average distance to the real values but in absolute units, the situation is the same as in the previous errors. The model pre-trained with correct simulated data yields the lowest value (3.71 kW ±1.18), followed by the model without pre-training (4.17 kW ± 1.03) and the model with a random pre-training (6.37 % ± 1.96).

TABLE IV. Average Results of a Cross-Validation Experiment Considering an Expanding Window and Considering the Accuracy of the Models. The Average CV(RMSE), the Average NMBE and the Average MAE Are Presented Together With Their Standard Deviations (SD)

Pre-training	CV(RMSE) [%]	SD	NMBE [%]	SD	MAE [kW]	SD
Correct	12.84	3.22	0.07	0.04	3.71	1.18
Random	21.15	5.87	0.10	0.08	6.37	1.96
None	14.22	3.21	0.06	0.04	4.17	1.03

Moreover, Fig. 5 shows the performance of the three built models over an entire week (specifically, from 12 July 2022 to17 July 2022) and considering the similar training process used for the previous accuracy analysis. It can be seen that, as presented in Table IV and considering the CV(RMSE) results, the model with correct pre-training best replicate the real behaviour of the studied PV installation. Although the model with no pre-training exhibits an accuracy not too far from the model just mentioned, the model pre-trained with incorrect simulated data is far from the real data.

In short, it is demonstrated from a speed and accuracy approach that pre-training the deep learning model with synthetic data is an effective way to improve its performance. Furthermore, in order to efficiently get this improvement, it is important to use correct simulated data. Pre-training with appropriate synthetic data allows to reduce the number of epochs, and thus the computational time, required in the training process by more than six times compared with no pre-training (see Table III). However, the use of incorrect simulated data, although faster than the model without pre-training, increases the computational time required in training by more than three times compared with the model correctly pre-trained. With respect to the accuracy of the models based on similar trainings, the use of incorrect simulated data, again, generates a model significantly less accurate than the model with a correct pre-training, but also than the model without pre-training (see Table IV). In this case, focusing only on the final average errors, both the model with no pre-training and the model with the correct pre-training show a similar performance, although the model pre-trained with the correct synthetic data achieved lower errors. These results show that although pre-training with synthetic data can provide more speed adapting to real data, if the data used is not appropriate, the accuracy of the model can stagnate and not reach the levels that would be achieved without pre-training.

Comparing the results of the proposed research with previous similar studies, taking into account the differences between installations and the improvements shown by the pre-trainings, the built models show error values lower or in the same range [59]–[61] and complying with the ASHRAE Guidelines [61], [62].

V. Conclusions

A methodology for optimising deep learning model configurations and improving their performance by means of pre-training based on synthetic data is presented in this paper. In this way, a great time reduction can be obtained, not only considering the reduction in the

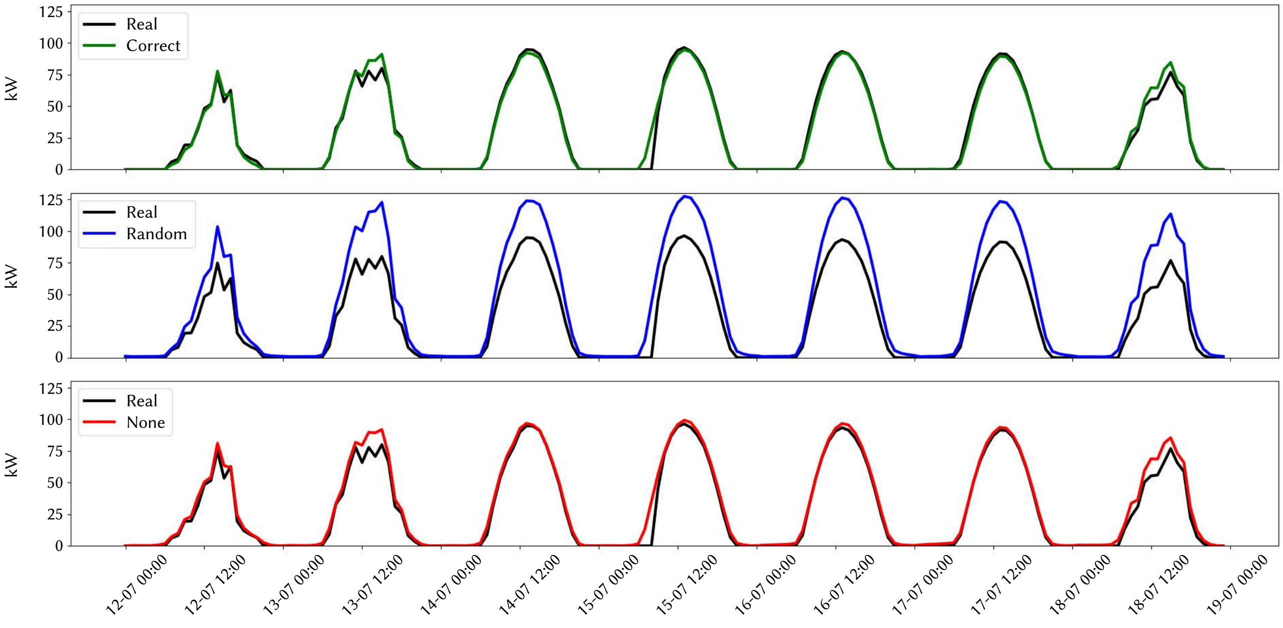


Fig. 5. Results of PV generation estimations in the second week of July 2022. All models built are considered for comparison. The CV(RMSE) values are 10:61 % for the model with correct pre-training, 47:63 % for the model with random pre-training and 13:31 % for the model without pre-training.

model training time but also in the time devoted to the search for the optimal architecture and the optimal training stopping criterion of the deep learning model. The study was conducted with an LSTM neural network built to perform PV generation predictions and pre-trained using synthetic data acquired with TRNSYS software.

On the one hand, the results achieved demonstrate that it is possible to pre-train a deep learning model, both with data simulated using the correct parameters and using random parameters, significantly reducing the time (measured in epochs and seconds) spent in the training process. The computational time required for the model to reach specific training error values is reduced by up to six times. In addition, in relation to the optimisation of DL model configurations, the proposed method (based on a multiobjective genetic algorithm) also reduces to less than half the evaluations needed to search all possible configurations and select an optimal one. On the other hand, the impact of using synthetic data generated with erroneous parameters is also analysed. In this case, an inadequate pre-training not only does not come close to the performance of a correct pre-training, but even can worsens the situation without pre-training. With regard to the accuracy of the built models considering the same training process on real data, it is shown that an incorrect pre-training produces less accurate models when fed with real data than a correct pre-training or a model without pre-training. The former two show a similar final accuracy but the model pre-trained with data simulated considering the correct parameters yields lower average errors. Here is a key insight of the research: although a pre-training with synthetic data may provide higher speed of adaptation to reality, if the data used in this pre-trainig are far from the real situation, it will affect to the final accuracy of the model and even lead to a worse performance compared to a model without pre-training.

The main limitation of this research is the amount of data. The monitoring period could be longer to reach a full year and the availability of data from more PV installations would make this study more consistent. The main outcome of this study is the evidence that the presented methodology can contribute to improve the performance of deep learning models. First, the multiobjective genetic algorithm NSGA-II allows us to use an efficiently optimised LSTM neural network without the need to evaluate all possible hyperparameter options. Second, the use of synthetic data to pre-train the built model allows us to significantly reduce the time spent on training and even slightly improve the final accuracy of the model. In this way, these improvements can be focused on making the use and distribution of photovoltaic energy more efficient. Thus, the fulfilment of the European Commission targets, commented at the beginning of the paper, will be closer. Lastly, this research evidences the importance of selecting adequate datasets for pre-training and generating global models that, once trained with simulated data, are used in real PV installations.

As future lines of research, more installations based on different parameters and different deep learning models could be considered to develop a more complete comparison and analysis. Using more installations to pre-train the deep learning model, or plug it in with a model that estimates the correct installation parameters from monitored data, can generate a global model that can be applied to different installations instead of having to pre-train the model with data specific to the particular installation under study.

Acknowledgments

This research was supported by the Ministry of Science, Innovation and Universities of the Spanish government under the DEEPSMART project (PID2021-126739OB-C21). The authors also want to thank the Ministry of Science, Innovation and Universities for grant FPU19/01187. Funding for open access charge: Universidade de Vigo/CISUG.

References

[1] A. Abdelkader, A. Rabeh, D. Mohamed Ali, J. Mohamed, "Multi-objective genetic algorithm based sizing optimization of a stand-alone wind/pv power supply system with enhanced battery/supercapacitor hybrid energy storage," *Energy*, vol. 163, pp. 351–363, 2018, doi: https://doi.org/10.1016/j.energy.2018.08.135.

[2] J. Munkhammar, J. D.K. Bishop, J. J. Sarralde, W. Tian, R. Choudhary, "Household electricity use, electric vehicle home-charging and distributed photovoltaic power production in the city of westminster," *Energy and Buildings*, vol. 86, pp. 439–448, 2015, doi: https://doi.org/10.1016/j.enbuild.2014.10.006.

[3] R. Fachrizal, M. Shepero, D. Van der Meer, J. Munkhammar, J. Widén, "Smart charging of electric vehicles considering photovoltaic power production and electricity consumption: A review," *eTransportation*, vol. 4, p. 100056, 2020, doi: https://doi.org/10.1016/j.etran.2020.100056.

[4] D. B. Richardson, "Electric vehicles and the electric grid: A review of modeling approaches, impacts, and renewable energy integration," *Renewable and Sustainable Energy Reviews*, vol. 19, pp. 247–254, 2013, doi: https://doi.org/10.1016/j.rser.2012.11.042.

[5] A. Anand, A. Shukla, H. Panchal, A. Sharma, "Thermal regulation of photovoltaic system for enhanced power production: A review," *Journal of Energy Storage*, vol. 35, p. 102236, 2021, doi: https://doi.org/10.1016/j.est.2021.102236.

[6] A. R. Jordehi, "Parameter estimation of solar photovoltaic (pv) cells: A review," *Renewable and Sustainable Energy Reviews*, vol. 61, pp. 354–371, 2016, doi: https://doi.org/10.1016/j.rser.2016.03.049.

[7] S. Theocharides, G. Makrides, A. Livera, M. Theristis, P. Kaimakis, G. E. Georghiou, "Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing," *Applied Energy*, vol. 268, p. 115023, 2020, doi: https://doi.org/10.1016/j.apenergy.2020.115023.

[8] S. Dwyer, S. Teske, "Renewables 2018 global status report," *Renewables 2018 Global Status Report*, 2018.

[9] I. Renewable Energy Statistics, "International renewable energy agency," *Renewable Energy Target Setting, Abu Dhabi, UAE*, 2015.

[10] E. Bueno, P. d. S. Vicente, T. C. Pimenta, E. R. Ribeiro, "Photovoltaic array reconfiguration strategy for maximization of energy production," *International Journal of Photoenergy*, vol. 2015, p. 592383, 2015, doi: 10.1155/2015/592383.

[11] L. Prado Osco, J. Marcato Junior, A. P. Marques Ramos, L. A. de Castro Jorge, S. Narges Fatholahi, J. de Andrade Silva, E. T. Matsubara, H. Pistori, W. N. Gonçalves, J. Li, "A review on deep learning in uav remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 102, p. 102456, 2021, doi: https://doi.org/10.1016/j.jag.2021.102456.

[12] M. Martínez-Comesaña, L. Febrero-Garrido, E. Granada-Álvarez, J. Martínez-Torres, S. Martínez-Mariño, "Heat loss coefficient estimation applied to existing buildings through machine learning models," *Applied Sciences*, vol. 10, no. 24, 2020, doi: 10.3390/app10248968.

[13] M. Nabipour, P. Nayyeri, H. Jabani, S. S., A. Mosavi, "Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis," *IEEE Access*, vol. 8, pp. 150199–150212, 2020, doi: 10.1109/ACCESS.2020.3015966.

[14] L. Mert, "Agnostic deep neural network approach to the estimation of hydrogen production for solar-powered systems," *International Journal of Hydrogen Energy*, vol. 46, no. 9, pp. 6272–6285, 2021, doi: https://doi.org/10.1016/j.ijhydene.2020.11.161.

[15] P. Dhanith, B. Surendiran, S. Raja, "A word embedding based approach for focused web crawling using the recurrent neural network," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 122-132, 2021.

[16] M. Martínez-Comesaña, L. Febrero-Garrido, F. Troncoso-Pastoriza, J. Martínez-Torres, "Prediction of building's thermal performance using lstm and mlp neural networks," *Applied Sciences*, vol. 10, no. 21, 2020, doi: 10.3390/app10217439.

[17] R. R. Kumari, V. V. Kumar, K. R. Naidu, "Optimized dwt based digital image watermarking and extraction using rnn-lstm," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 2, pp. 150-162, 2021.

[18] X.-H. Le, H. V. Ho, G. Lee, S. Jung, "Application of long short-term

memory (lstm) neural network for flood forecasting," *Water*, vol. 11, no. 7, 2019, doi: 10.3390/w11071387.

[19] M. Chai, F. Xia, S. Hao, D. Peng, C. Cui, W. Liu, "Pv power prediction based on lstm with adaptive hyperparameter adjustment," *IEEE Access*, vol. 7, pp. 115473–115486, 2019, doi: 10.1109/ACCESS.2019.2936597.

[20] M. Khodayar, M. E. Khodayar, S. Mohammad Jafar Jalali, "Deep learning for pattern recognition of photovoltaic energy generation," *The Electricity Journal*, vol. 34, no. 1, p. 106882, 2021, doi: https://doi.org/10.1016/j.tej.2020.106882.

[21] M. S. Hossain, H. Mahmood, "Short-term photovoltaic power forecasting using an lstm neural network and synthetic weather forecast," *IEEE Access*, vol. 8, pp. 172524–172533, 2020, doi: 10.1109/ACCESS.2020.3024901.

[22] M. Abdel-Nasser, K. Mahmoud, "Accurate photovoltaic power forecasting models using deep lstm-rnn," *Neural Computing and Applications*, vol. 31, no. 7, pp. 2727–2740, 2019, doi: 10.1007/s00521-017-3225-z.

[23] L. Yang, A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020, doi: https://doi.org/10.1016/j.neucom.2020.07.061.

[24] M. Martínez-Comesaña, P. Eguía-Oller, J. Martínez-Torres, L. Febrero-Garrido, E. Granada-Álvarez, "Optimisation of thermal comfort and indoor air quality estimations applied to in-use buildings combining nsga-iii and xgboost," *Sustainable Cities and Society*, vol. 80, p. 103723, 2022, doi: https://doi.org/10.1016/j.scs.2022.103723.

[25] Y. Yoo, "Hyperparameter optimization of deep neural network using univariate dynamic encoding algorithm for searches," *Knowledge-Based Systems*, vol. 178, pp. 74–83, 2019, doi: https://doi.org/10.1016/j.knosys.2019.04.019.

[26] Y. Novaria Kunang, S. Nurmaini, D. Stiawan, B. Yudho Suprapto, "Attack classification of an intrusion detection system using deep learning and hyperparameter optimization," *Journal of Information Security and Applications*, vol. 58, p. 102804, 2021, doi: https://doi.org/10.1016/j.jisa.2021.102804.

[27] Y. Guo, J.-Y. Li, Z.-H. Zhan, "Efficient hyperparameter optimization for convolution neural networks in deep learning: A distributed particle swarm optimization approach," *Cybernetics and Systems*, vol. 52, no. 1, pp. 36–57, 2021, doi: 10.1080/01969722.2020.1827797.

[28] S. Lee, J. Kim, H. Kang, D.-Y. Kang, J. Park, "Genetic algorithm based deep learning neural network structure and hyperparameter optimization," *Applied Sciences*, vol. 11, no. 2, 2021, doi: 10.3390/app11020744.

[29] H. Chung, K.-S. Shin, "Genetic algorithm-optimized multi-channel convolutional neural network for stock market prediction," *Neural Computing and Applications*, vol. 32, no. 12, pp. 7897–7914, 2020, doi: 10.1007/s00521-019-04236-3.

[30] Á. A. Domingo, M. A. Ezquerra, "Gpgpu implementation of a genetic algorithm for stereo refinement," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 3, no. 2, pp. 69–76, 2015.

[31] M. Martínez-Comesaña, A. Ogando-Martínez, F. Troncoso-Pastoriza, J. López-Gómez, L. Febrero-Garrido, E. Granada-Álvarez, "Use of optimised mlp neural networks for spatiotemporal estimation of indoor environmental conditions of existing buildings," *Building and Environment*, vol. 205, p. 108243, 2021, doi: https://doi.org/10.1016/j.buildenv.2021.108243.

[32] R. Damaševičius, M. Gupta, N. Kumar, B. K. Singh, N. Gupta, "Nsga-iii-based deep-learning model for biomedical search engines," *Mathematical Problems in Engineering*, vol. 2021, p. 9935862, 2021, doi: 10.1155/2021/9935862.

[33] P. Trampert, D. Rubinstein, F. Boughorbel, C. Schlinkmann, M. Luschkova, P. Slusallek, T. Dahmen, S. Sandfeld, "Deep neural networks for analysis of microscopy images—synthetic data generation and adaptive sampling," *Crystals*, vol. 11, no. 3, 2021, doi: 10.3390/cryst11030258.

[34] J. Liu, J. Gu, H. Li, K. H. Carlson, "Machine learning and transport simulations for groundwater anomaly detection," *Journal of Computational and Applied Mathematics*, vol. 380, p. 112982, 2020, doi: https://doi.org/10.1016/j.cam.2020.112982.

[35] K. Antczak, "Deep recurrent neural networks for ecg signal denoising," 2018. [Online]. Available: https://arxiv.org/abs/1807.11551, doi: 10.48550/ARXIV.1807.11551.

[36] Q. Wang, J. Gao, W. Lin, Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[37] J. C. Balloch, I. Agrawal, Varun Essa, S. Chernova, "Unbiasing semantic segmentation for robot perception using synthetic data feature transfer," 2018. [Online]. Available: https://arxiv.org/abs/1809.03676, doi: 10.48550/ARXIV.1809.03676.

[38] Y. Yu, X. Si, C. Hu, J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[39] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020, doi: https://doi.org/10.1016/j.physd.2019.132306.

[40] Y. Hu, A. E. G. Huber, J. Anumula, S. Chii Liu, "Overcoming the vanishing gradient problem in plain recurrent networks," *CoRR*, vol. abs/1801.06105, 2018.

[41] S. Jha, A. Dey, R. Kumar, V. Kumar, "A novel approach on visual question answering by parameter prediction using faster region based convolutional neural network," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 5, pp. 30–37, 2019.

[42] A. H. Elkasem, S. Kamel, A. Rashad, F. J. Melguizo, "Optimal performance of doubly fed induction generator wind farm using multi-objective genetic algorithm," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 5, pp. 48–53, 2019.

[43] S. Katoch, S. S. Chauhan, V. Kumar, "A review on genetic algorithm: past, present, and future," *Multimedia Tools and Applications*, vol. 80, no. 5, pp. 8091–8126, 2021, doi: 10.1007/s11042-020-10139-6.

[44] A. Vukadinović, J. Radosavljević, A. Đorđević, M. Protić, N. Petrović, "Multi-objective optimization of energy performance for a detached residential building with a sunspace using the nsga-ii genetic algorithm," *Solar Energy*, vol. 224, pp. 1426–1444, 2021, doi: https://doi.org/10.1016/j.solener.2021.06.082.

[45] S. Wang, D. Zhao, J. Yuan, H. Li, Y. Gao, "Application of nsga-ii algorithm for fault diagnosis in power system," *Electric Power Systems Research*, vol. 175, p. 105893, 2019, doi: https://doi.org/10.1016/j.epsr.2019.105893.

[46] M. A. J. Idrissi, H. Ramchoun, Y. Ghanou, M. Ettaouil, "Genetic algorithm for neural network architecture optimization," in *2016 3rd International Conference on Logistics Operations Management (GOL)*, 2016, pp. 1–4.

[47] Y. Wu, J. Wei, W. Ying, Y. Lan, Z. Cui, Z. Wang, "A collaborative decomposition-based evolutionary algorithm integrating normal and penalty-based boundary intersection methods for many-objective optimization," *Information Sciences*, vol. 616, pp. 505–525, 2022, doi: https://doi.org/10.1016/j.ins.2022.10.136.

[48] S. Verma, M. Pant, V. Snasel, "A comprehensive review on nsga-ii for multi-objective combinatorial optimization problems," *IEEE Access*, vol. 9, pp. 57757–57791, 2021, doi: 10.1109/ACCESS.2021.3070634.

[49] F. Troncoso-Pastoriza, M. Martínez-Comesaña, A. Ogando-Martínez, J. López-Gómez, P. Eguía-Oller, L. Febrero-Garrido, "Iot-based platform for automated ieq spatio-temporal analysis in buildings using machine learning techniques," *Automation in Construction*, vol. 139, p. 104261, 2022, doi: https://doi.org/10.1016/j.autcon.2022.104261.

[50] Z. Pang, F. Niu, Z. O'Neill, "Solar radiation prediction using recurrent neural network and artificial neural network: A case study with comparisons," *Renewable Energy*, vol. 156, pp. 279–289, 2020, doi: https://doi.org/10.1016/j.renene.2020.04.042.

[51] U. O. W.-M. Solar Energy Laboratory, *TRNSYS, a transient simulation program*. Madison, Wis. : The Laboratory, 1975., 1975.

[52] A. Remlaoui, D. Nehari, M. Laissaoui, A. M. Sandid, "Performance evaluation of a solar thermal and photovoltaic hybrid system powering a direct contact membrane distillation: Trnsys simulation," *Desalin. Water Treat*, vol. 194, pp. 37–51, 2020.

[53] J. López Gómez, A. Ogando Martínez, F. Troncoso Pastoriza, L. Febrero Garrido, E. Granada Álvarez, J. A. Orosa García, "Photovoltaic power prediction using artificial neural networks and numerical weather data," *Sustainability*, vol. 12, no. 24, 2020, doi: 10.3390/su122410295.

[54] X. de Galicia, "Meteogalicia," 2021. [Online]. Available: https://www.meteogalicia.gal/web/RSS/ rssIndex.action, Last access: 12 September 2022.

[55] NOAA, "The global forecast system (gfs) - global spectral model (gsm)," 2021. [Online]. Available: https://www. emc.ncep.noaa.gov/emc/pages/numerical_forecast_systems/gfs/documentation.php, Last access: 12 September 2022.

[56] G. Van Rossum, F. L. Drake Jr, "Python/c api reference manual," *Python Software Foundation*, 2002.

[57] S. Martínez, E. Pérez, P. Eguía, A. Erkoreka, E. Granada, "Model calibration and exergoeconomic optimization with nsga-ii applied to a residential cogeneration," *Applied Thermal Engineering*, vol. 169, p. 114916, 2020, doi: https://doi.org/10.1016/j.applthermaleng.2020.114916.

[58] A. E. I. Brownlee, J. A. Wright, "Constrained, mixed-integer and multi-objective optimisation of building designs by nsga-ii with fitness approximation," *Applied Soft Computing*, vol. 33, pp. 114–126, 2015, doi: https://doi.org/10.1016/j.asoc.2015.04.010.

[59] M. K. Park, J. M. Lee, W. H. Kang, J. M. Choi, K. H. Lee, "Predictive model for pv power generation using rnn (lstm)," *Journal of Mechanical Science and Technology*, vol. 35, no. 2, pp. 795–803, 2021, doi: 10.1007/s12206-021-0140-0.

[60] B. Kim, D. Suh, "Solar photovoltaic generation forecasting using machine learning methods," *The Journal of Contents Computing*, vol. 2, no. 1, pp. 105–112, 2020.

[61] B. Kim, D. Suh, "A hybrid spatio-temporal prediction model for solar photovoltaic generation using numerical weather data and satellite images," *Remote Sensing*, vol. 12, no. 22, 2020, doi: 10.3390/rs12223706.

[62] ANSI/ASHRAE, "Measurement of energy and demand savings," in *ASHRAE Guideline 14-2002*, vol. 8400, 2002, p. 170.

Miguel Martínez-Comesaña

Miguel Martínez Comesaña was born in Vigo (Spain). He holds a degree in Economics since 2017. He obtained his Master in Statistics from the University of Santiago de Compostela in 2019. He received his PhD in Artificial Intelligence from the University of Vigo in 2023. He is the author of several articles specialized in the application of AI in energy efficiency analysis. Currently, he is Data Scientist and Engineer at the University of Vigo, being part of the research group GTE (Energy Technology Group).

Javier Martínez Torres

Javier Martínez Torres is a Mathematician and Engineering PhD from the University of Vigo. He is currently an Assistant Professor at the University of Vigo and has participated in more than 20 research projects as principal investigator. He has published more than 60 papers in JCR indexed journals and participate in more than 35 international conferences.

Pablo Eguía Oller

Pablo Eguía Oller is an engineer and Engineering PhD from the University of Vigo. Main researcher in the Energy Technology research group (GTE) and has been working for 10 years on indoor air quality and energy efficiency in buildings in both European and national projects. More than 70 papers in JCR indexed journals.

Javier López-Gómez

Javier López-Gómez holds a PhD in Energy Efficiency from the University of Vigo. Specialized in the collection, processing and analysis of data applied to multiple fields of engineering (energy generation and consumption, meteorological phenomena, forest fires, environmental quality in buildings, or coastal marine flows).

Regular Issue

Stacked LSTM for Short-Term Wind Power Forecasting Using Multivariate Time Series Data

Manisha Galphade[1,2]*, V. B. Nikam[1], Biplab Banerjee[3], Arvind W. Kiwelekar[4], Priyanka Sharma[5]

[1] Department of Computer Engineering & IT, Veermata Jijabai Technological Institute, Mumbai (India)
[2] School of Computing, MIT Art Design & Technology University, Pune (India)
[3] Center of Studies in Resources Engineering, IIT Bombay Mumbai, (India)
[4] Department of Information Technology, Dr Babasaheb Ambedkar Technological University, Lonere, Raigad (India)
[5] Director of Software Engineering (HPC AI R&D Lab), Fujitsu Research (FRIPL),AI Thought Leader, Bengaluru, Karnataka (India)

* Corresponding author: galphademanisha@gmail.com

Received 20 May 2023 | Accepted 10 May 2024 | Published 15 July 2024

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

Currently, wind power is the fast growing area in the domain of renewable energy generation. Accurate prediction of wind power output in wind farms is crucial for addressing the challenges associated the power grid. This precise forecasting enables grid operators to enhance safety and optimize grid operations by effectively managing fluctuations in power generation, ensuring a reliable and stable energy supply. In recent years, there has been a significant rise in research and investigations conducted in this field. This study aims to develop a multivariate short-term wind power forecasting (WPF) model with the objective of enhancing forecasting precision. Among the various prediction models, deep learning models such as Long Short-Term Memory (LSTM) have demonstrated outstanding performance in the field of WPF. By adding multiple layers of LSTM networks, the model can capture more complex patterns. To improve the performance, data pre-processing is carried out using two techniques such as removal of missing values and imputing missing values using Random Forest Regressor (RFR). The comparison between the proposed Stacked LSTM model and other methods including vector autoregressive (VAR), Multiple Linear Regression, Gated Recurrent Unit (GRU) and Bidirectional LSTM (BiLSTM) has been experimented on two datasets. The experimental results show that after imputing missing values using RFR, the Stacked LSTM is optimized model for better performance than above mentioned reference models.

Keywords

Deep Learning, Long Short Term Memory, Multivariate, Renewable Energy , Time Series Data, Wind Power Forecasting.

DOI: 10.9781/ijimai.2024.07.002

I. Introduction

Renewable energy sources are developed and adopted since there is a decline in the cost of renewable technology, globally. This is for the reason that using conventional energy causes the environment irreparable harm [1]. Among them, wind power is an attractive option because of its low operating costs, low environmental impact, and high availability and sustainability. As a result, to assure the stability and safety of the power system, efficient and reliable wind power forecasting (WPF) methodologies are required. The three categories of WPF tasks are: (i) Short-Term Prediction: time range from several minutes to several hours; (ii) Medium-Term Prediction: time range from hours to a week; and (iii) Long-Term Prediction: time range from week to year or more.Short-term WPF allows the power industry to prepare for fluctuations in wind farm output. This preparation reduces operating expenses, the need for backup power, and reduce the strain on the electrical grid. The relevant methods mainly classified into physical, statistical, and hybrid methods [2].

The physical method [3] requires lots of data, because the establishment of a predictive model will be done with the help of many variables, and the amount of data has a direct relationship with forecast accuracy. As a result, the predictive model's calculation process and mathematical structure are complex, resulting in a relatively longer computation time. The most extensively utilized method is numerical weather prediction (NWP) [4]. The NWP involves extensive calculations and is better suited for Long-Term forecasting rather than Medium-Term and Short-Term forecasting. Statistical methods involve linear models and non-linear models. Autoregressive (AR), Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA) [5], vector autoregressive (VAR), Smooth-Transition Auto-Regressive (STAR) are some examples of linear

Please cite this article in press as:
M. Galphade, V. B. Nikam, B. Banerjee, A. W. Kiwelekar, P. Sharma. Stacked LSTM for Short-Term Wind Power Forecasting Using Multivariate Time Series Data, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 71-81, 2025, http://dx.doi.org/10.9781/ijimai.2024.07.002

model. ARIMA is unable to reliably predict wind power due to the accumulation of error hence many strategies have been developed to solve this limitation. The Non-linear model, includes strategies such as Autoregressive Conditional Heteroskedasticity (ARCH) [6], Generalized Autoregressive Conditional Heteroskedasticity Process (GARCH) and Deep learning algorithms [7], [8]. Artificial intelligence is a recent area of statistical methods, which is successfully applied for forecasting wind power, and the results of the forecasts gaining people approval [9]–[14]. Techniques used here involve artificial neural network (ANN) [15], support vector machine (SVM) [16], and Long Short-Term Memory (LSTM) [17]. Artificial intelligence is best suited for general purposes, but its disadvantage is that it cannot accurately explain the relationship between model elements. A study conducted by Kia Qu et al. investigated the use of the transformer architecture for short-term WPF [18]. This research covered the examination of wind power output data from various wind farms and included a comparative assessment with the LSTM model serving as a benchmark model. Where as a novel deep learning model that combines transformers and wavelet transforms uses weather data to predict wind speed and power generation six hours in advance, outperforming LSTM models [19]. Transformers offer several advantages for WPF, they come with computational requirements and may require large datasets for training. Additionally, model architecture, hyperparameter tuning, and data preprocessing are critical factors in achieving accurate forecasts.

A hybrid model [20] is a combination of different forecasting models. These combined models are expected to avoid the shortcomings of one model in forecasting wind power. Statistical methods are best in achieving high accuracy for short-term forecasts but tend to accumulate errors when used for long-term forecasts. On the other hand, physical methods, owing to their extensive scope and scale, are better suited for long-term forecasting rather than short-term predictions. Therefore, there are many errors in the existing methods, and studies are currently underway to improve the forecasting methods. Hybrid methods are generally classified into four types: hybrid approach based on data pre-processing technique, weight-based hybrid approach, hybrid approach based on data post-processing technique and parameter selection and optimization technique [21]. Table I provides a brief assessment of each class of hybrid techniques. This research takes first type of hybrid approach. Recent studies have suggested large amount of hybrid techniques in order to decrease the prediction errors by incorporating the merits of different methods. The author E. Lopez et al. utilized Principal Component Analysis (PCA) to reduce the input variables of the LSTM model for NWP data prediction [22]. The proposed model has demonstrated superior accuracy in comparison to SVM and Backpropagation Neural Network (BPNN). LSTM has higher prediction accuracy; therefore, it is used in wider range of applications. The author F. Shahid et al. merged wavenets with LSTM for diminishing wave and gradient transformation for nonlinear mapping [23]. WN-LSTM is applied on seven wind farms in Europe for Short-Term WPF. The performance is assessed based on metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The results indicate a notable improvement, with performance showing a significant increase of 30%. In the Advanced LSTM [24] technique, the author fine-tuned the neural network's parameters based on new insights gained over time, instead of retraining the model using the entire dataset. The results of a case study in Belgium show that recalibration of advanced models will improve the consistency of predictions while reducing the cost of evaluating the system. Z. Sun et al. introduced a hybrid model that combines variational mode decomposition (VMD), Convolutional Long Short-Term Memory network (ConvLSTM), and error analysis [25]. The VMD approach divides the input wind energy into a set of different frequency components and then obtains a new structure that incorporates the convolutional layer into the LSTM network. This structure can extract the spatio-temporal features of each sub-series as an initial prediction engine.

Taking into account these issues, this study presents stacked LSTM for Short-Term Wind Power Forecasting using Multivariate Time Series Data. The major contributions of our paper are summarized as follows: (a) The presentation of a wind power forecasting model, employing the deep recurrent neural network LSTM, to assess potential improvements in predictions through the incorporation of additional training layers within the framework of time series analysis. (b) The proposed model uses data pre-processing technique and hyperparameter optimization to improve the LSTM network structure and proposed model is evaluated by comparing predicted values with actual values based on MAE & RMSE statistical error measures. (c) The performance of twelve deep LSTM models are assessed using various architectures for WPF. (d) The effectiveness of the presented research work is compared with three methods used in the literature. The subsequent sections of the paper are structured as follows: Section II describes overviews of LSTM. Section III provides an overview of the data pre-processing method. Section IV explains proposed model and performance metrics. Section V outlines experimental analysis on two open-access datasets. Finally, Section VI presents concluding remarks and future work.

II. Theoretical Framework

In this section LSTM is illustrated in detail. LSTM [26], a refined version of RNN, is designed for working with time series data. It's like a unique neural network that excels at making decisions based on data

TABLE I. A Brief Evaluation of the Hybrid Approaches

Hybrid WPF Approach	Advantages	Disadvantages
Hybrid approach based data pre-processing technique [27], [28]	• Higher performance compared to other approaches	• Mathematical knowledge of decomposition is required • slow response to new data
Weight-based hybrid approach [29]	• Adaptive to new data • Easy to implement • Adaptable to a variety of situations	• To determining the weights, an additional model is required • The best forecast within the forecast range is not guaranteed
Hybrid approach based on data post-processing technique [11]	• Effective in reducing systematic error • High accuracy	• Depends on designer's understanding of the optimization problems • Harder to code • Computationally complex
Hybrid approach based on parameter selection and optimization technique [30], [31]	• Relatively basic structure	• Computational time inefficient

over extended periods. Instead of RNN's conventional hidden layers, LSTM employs memory cells, as shown in the Fig. 1, to effectively manage and retain information.

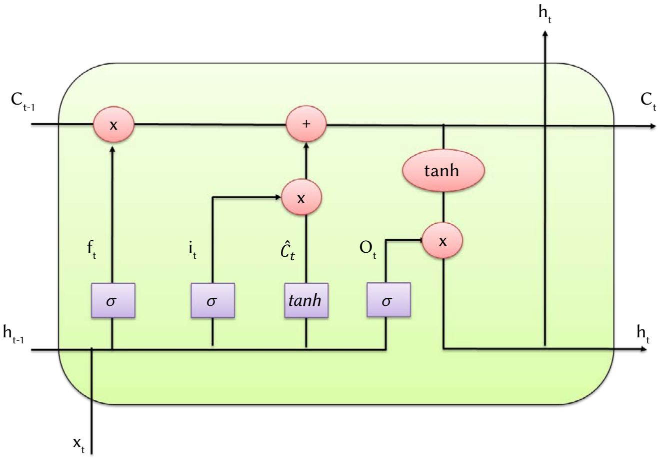


Fig. 1. Internal structure of LSTM cell.

LSTMs are experts at grasping temporal relationships by employing several gate units like the input gate (i_t), the forget gate (f_t), and the output gate (o_t), all accompanied by activation functions like tanh and sigmoid. By utilizing the sigmoid function each gate produces a binary output, either 1 or 0, which satisfies the input data flow. The input gate is a critical element responsible for identifying and selecting relevant information from the current input. It takes both the present input and the previous hidden state into account, using them to compute a candidate cell state. The decision related to the kind of information to be maintained and discarded is the responsibility of the forget gate. It achieves this by considering the current input and the previous hidden state, ultimately calculating a forget factor. The process of updating the cell state involves combining the previous cell state with the candidate cell state, with the assistance of the forget gate and input gate. This step ensures that the cell state is updated to integrate new information while preserving essential past information. The output gate plays a pivotal role in deciding which information from the updated cell state should be forwarded as the output. Additionally, it is instrumental in guiding subsequent cells in a sequence.

The mathematical computation is explained as follows:

Equation (1) compute the output vector h_t

$$\text{output: } h_t = o_t * \tanh{(C_t)} \tag{1}$$

where o_t is output gate vector, C_t is cell state, expressed using Equation (2) and Equation (3).

$$\text{OutputGate : } o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \tag{2}$$

$$\text{CellUpdate : } C_t = f_t * C_{t-1} + i_t * \hat{C}_t)) \tag{3}$$

To adjust the memory cell, at each step the hyperbolic tangent (tanh) function is utilized so as to improve the training performance which is shown in Equation (4).

$$\text{ProcessInput: } \hat{C}_t = \tanh{(W_c \cdot [h_{t-1}, x_t] + b_c)} \tag{4}$$

The values f_t and i_t correspond to the forget gate and input gate, respectively. These values are computed using Equation (5) and Equation (6).

$$\text{ForgetGate : } f_t = \sigma\,(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{5}$$

$$\text{InputGate : } i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{6}$$

Where W represents the weight matrix. Here subscript *o*, *c*, *i* and *f* denotes weight of each gate. LSTM networks have established themselves as exceptionally effective technique across an array of applications which includes sequential data and time series analysis. Their versatility is well-demonstrated in various domains. LSTM plays a crucial role in the domain of Natural Language Processing (NLP) which includes machine translation using sequence-to-sequence analysis and Google Neural Machine Translation (GNMT) models [32]. Moreover, LSTMs are also used for text generation tasks, for instance OpenAI's GPT-2 model to craft human-like text [33]. LSTM (Long Short-Term Memory) networks are versatile in time series forecasting, proving effective in applications such as predicting stock prices [34], [35], oil production [36], weather forecasting [37], and Automatic Speech Recognition systems [38] ,Sea Surface Temperature Prediction [39], photovoltaic Generation [40] by analyzing historical data. They proficiently translate spoken language into text, enabling voice-controlled personal assistants like Siri, Google Assistant, and Amazon Alexa, accurately. In Healthcare, LSTMs are crucial for analyzing and decision making in medical time series data such as electrocardiogram (ECG) and electroencephalogram (EEG) [41]. In autonomous driving, LSTM networks are effectively employed for tasks like vehicle trajectory prediction and object detection [42]. Human Activity Recognition (HAR) utilizes LSTM networks in order to recognize and classify human activities emerging from sensor data located in wearable devices [43]. These examples collectively highlight the adaptability of LSTM networks in handling sequential data across a diverse range of domains.

III. Data Pre-Processing

Time series information is generally obtained from real world environment or data generated from sensors. These data are usually affected by instances like noise. A proper method for collecting data and the removal of any deficient information should be followed before performing any kind of analysis. This is crucial as it will assist in making the analysis less challenging and to avoid incorrect analytical conclusions. Typically, data obtained from wind farms contains a lot of missing values due to sensors malfunctioning. Detecting outliers, eliminating noise, and filling in missing values are the three most common data pre-processing techniques.

- **Detecting outliers**: Statistical-based outlier detection methods leverage statistical distributions and the relationship of data points to these distributions. Parametric methods [44], [45] rely on predefined models, while non-parametric methods take a more flexible and data-driven approach to identify outliers. Whereas density-based [46] outlier detection methods are founded on a fundamental principle: outliers tend to reside in regions of low data density. Conversely, non-outliers, often referred to as inliners or genuine data points, are expected to cluster in dense neighborhoods. These methods assess the densities of data points within their local contexts, comparing them to the densities of their nearby neighbors. Distance-based methods identify outliers by calculating distances between data points. Outliers are often defined as data points that are significantly distant from their nearest neighbors.
- **Eliminating noise**: One of the dominant challenges encountered in wind farm datasets is the presence of noisy data. This issue can be effectively addressed by employing a filter method to extract the accurate signal estimation. Among the frequently employed data filtering techniques, the most common strategies include frequency domain, time-domain and time-frequency domain filtering.
- **Missing values**: Missing data are common in statistical analyses. The imputation function is widely used to identify the missing values in a wind farm based on the variables and their relationship. Do not use metaphorical expressions.

The data was normalized before using it in the forecasting system. A dataset can be normalized using a variety of methods. The method used in this study is known as min-max normalization.

A. Sliding Window

In a multivariate time series, each variable represents an aspect of a complex system, and the temporal evolution of these variables signifies changes in the underlying system state. To capture the interrelationships and temporal dependencies among these variables and derive the system's state over specific periods, we employed a sliding window approach with a size of K to segment the multivariate time series data. This partitioning technique allows us to analyze the data within discrete intervals, facilitating the extraction of meaningful insights about the system's behavior. As illustrated in Fig. 2, for a given point in time t, we can establish the sliding window data $W_t \in R^{MXK}$, where $W_t = \{x_{t-K}, ..., x_{t-1}, x_t\}$, M is the dimension of the variable and K represent the sliding window size.

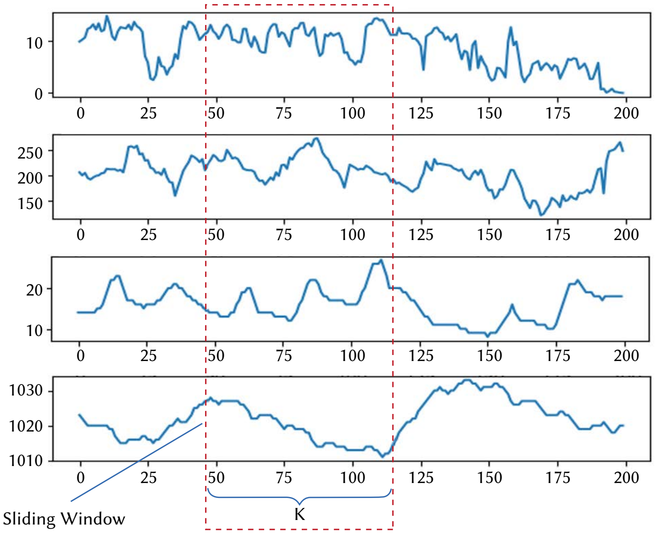


Fig. 2. Sliding Window.

B. RFR Based Imputation Method

Random forest [47] imputation is a machine learning technique that offers the advantage of handling nonlinearities and interactions within data without necessitating the specification of a specific regression model. These methods do not assume normality or require specification of parametric models. In this paper random forest method has been used for filling missing value. It extends from classification and regression trees (CART) [48], which are predictive models that iteratively partition the data based on predictor variable values. Notably, random forests don't depend on distributional assumptions and can effectively manage nonlinear relationships and interactions in the data. The schematic of the Random Forest Imputation framework is shown in Fig. 3 The steps involved in the proposed imputation approach is explained in Algorithm 1:

Algorithm 1: Random Forest Imputation Algorithm

Data: Data set $\mathcal{D}$ containing $|\mathcal{R}|$ records and $|\mathcal{A}|$ attributes.

Result: An imputed dataset $\mathcal{D}_{\text{Final}}$ containing $|\mathcal{R}|$ records and $|\mathcal{A}|$ attributes.

```
for each record $R_i \in \mathcal{D}$ do
    if any attribute $A_k$ is missing then
        $\mathcal{D}_{\text{miss}} = \mathcal{D}_{\text{miss}} \cup R_i$;
    end
    else
        $\mathcal{D}_{\text{comp}} = \mathcal{D}_{\text{comp}} \cup R_i$;
    end
end
$rf_{\text{model}} = \text{TrainRandomForest}(\mathcal{D}_{\text{comp}})$;
while $\mathcal{D}_{\text{miss}} \neq \emptyset$ do
    $R_i \leftarrow$ Select an incomplete record from $\mathcal{D}_{\text{miss}}$;
    for $j = 1, ..., M$ do
        if $r_{ij} = NaN$ then
            $r_{ij} = \text{PredictRFR}(rf_{model}, R_i)$;
            $\mathcal{D}_{\text{compute}} = \mathcal{D}_{\text{compute}} \cup A_j$ ;
        end
    end
end
$\mathcal{D}_{\text{Final}} = \mathcal{D}_{\text{comp}} \cup \mathcal{D}_{\text{compute}}$;
return $\mathcal{D}_{\text{Final}}$
```

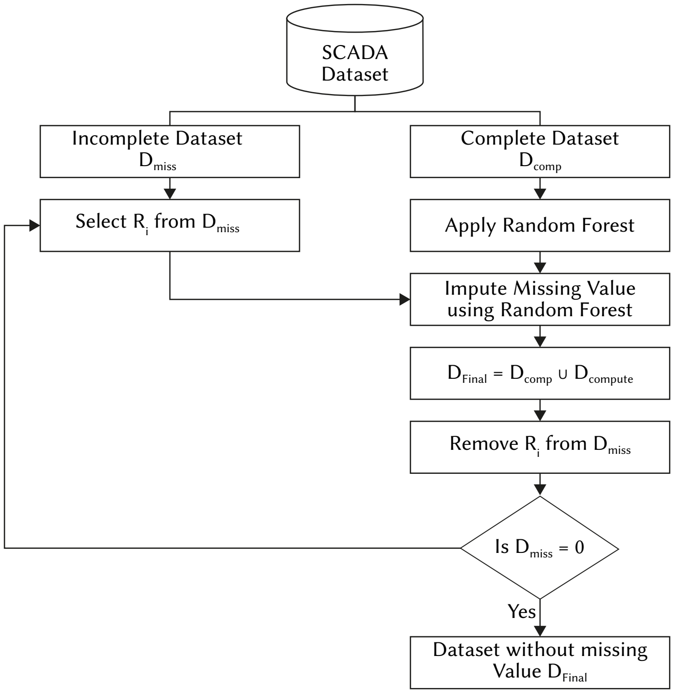

SCADA Dataset
Incomplete Dataset D_miss
Complete Dataset D_comp
Select R_i from D_miss
Apply Random Forest
Impute Missing Value using Random Forest
D_Final = D_comp ∪ D_compute
Remove R_i from D_miss
Is D_miss = 0
Yes
Dataset without missing Value D_Final

Fig. 3. RFR Imputation framework.

The effectiveness of the imputation method is evaluated by using RMSE and MAE parameters.

TABLE II. Details of Dataset Used in the Experiment

Data Source	Attribute Specifications	Resolution	Capacity	No. of records
Sotavento wind farm	Wind speed, wind direction and wind power	10-minute, 1-hour, 1-day, Data from 2014 to till date	17560kW	52568 records with 3 attributes
Yalova Wind Farm	Active power, Theoretical power, Wind speed, Wind direction	10 min period 1 January 2018 to 31 December 2018	54000kW	50530 records with 4 attribute

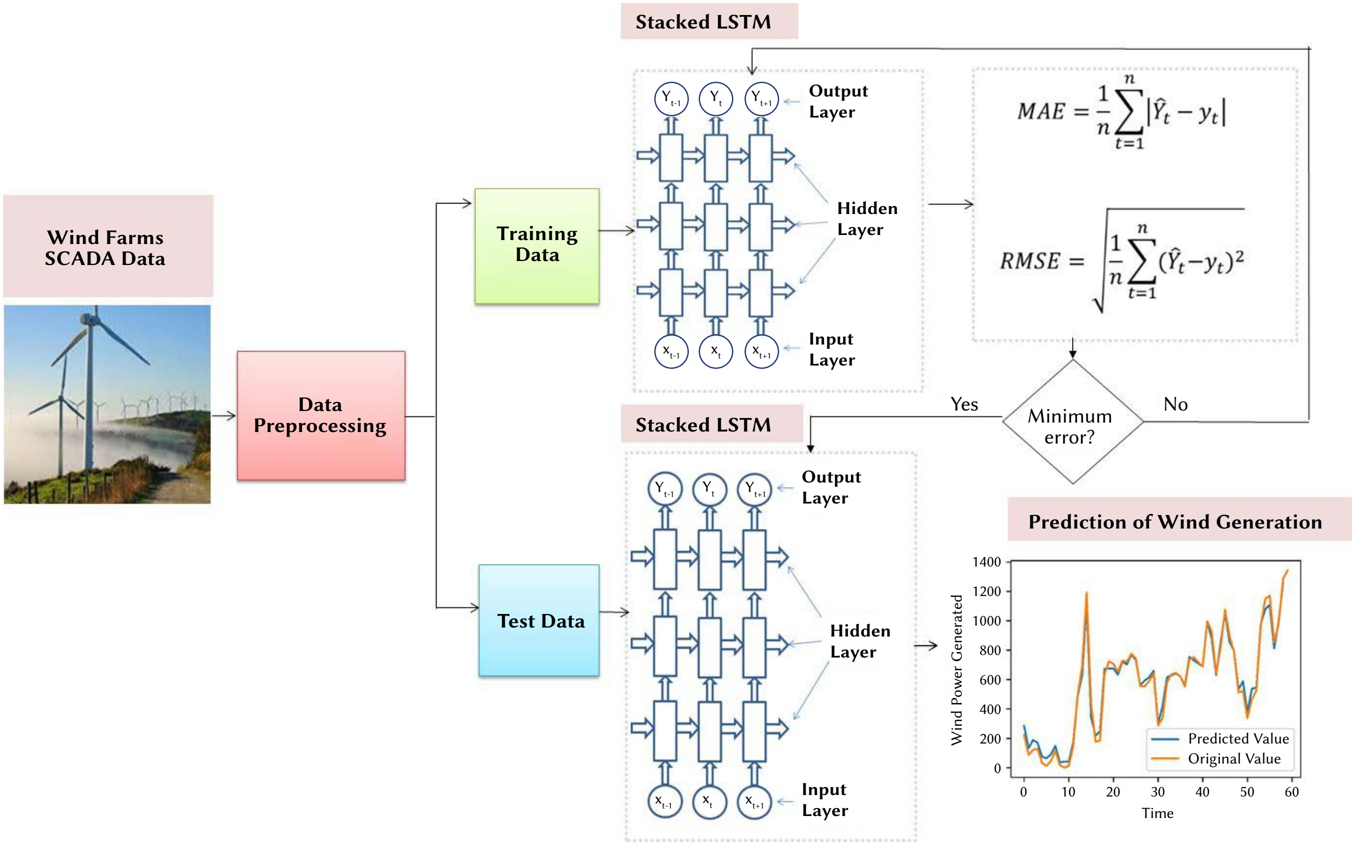


Fig. 4. Proposed stacked LSTM architecture.

C. Dataset

The characteristics of the two datasets include data instances, attributes, capacity, resolution, and record count which are detailed in Table II.

1. Sotavento Wind Farm

This dataset is from the Sotavento wind farm in Galicia, Spain with a total capacity of 17560 kW which consists of 24 onshore turbines. This wind farm gives real-time information on wind direction, wind speed and turbine's power generation. This paper uses five years data from 2014-2019 to test the proposed forecast model. The resolution of the data is hourly. Some attribute of this dataset has missing value. Wind power and wind speed is having 756 and 500 missing values respectively. Pre-processing strategies with deep learning methods are combined to evaluating the effectiveness of pre-processing. Two approaches have been used while performing experimentation, first is simply removing missing value and second approach is to use RFR imputer.

2. Yalova Wind Farm

The Yalova Wind Farm (YWF) situated in western Turkey records wind related information such as wind speed, wind direction, turbine power (theoretical) and generated. YWF possess a total capacity of 54,000 kW, obtained from 36 wind turbines. The wind generated information is collected and reserved utilizing a Supervisory Control and Data Acquisition (SCADA) system. Here the data is recorded at an interval of 10 minutes. The format of the data available is CSV.

IV. Proposed Model

Fig. 4 shows the proposed architecture. For the experimentation purpose two datasets are used, first is from Sotavento wind farm in Galicia and second from Yalova Wind in west Turkey. Data pre-processing is done on both the dataset as described in Section III. The dataset is partitioned into training and testing samples by employing 70:30 ratio. The model is trained on train set using stacked LSTM. In order to improve the accuracy of WPF, various network structures, along with different numbers of layers and neurons in each layer, are tested. Min-Max scaling function is employed to scale the input features before applying to the deep learning models. Min-Max scalar is expressed by Equation (7) as,

$$X_{\text{scaled}} = \frac{(X_i - \min(X))}{(\max(X) - \min(X))} \tag{7}$$

A. Stacked LSTM

A stacked LSTM is a neural network architecture that comprises multiple LSTM layers, creating a multi-layer structure, as illustrated in the Fig. 5. Incorporating multiple LSTM layers increases the model's complexity and depth. In this composite LSTM structure, each output layer of the LSTM model serves as input to subsequent layers within the same block. This design allows the model to capture the temporal patterns in the data and combine the learned representations from previous layers, resulting in a higher-level abstraction in the final output. Each intermediate LSTM layer produces a sequence vector as its output, which serves as the input for the next LSTM layer. Unlike a single-output LSTM, the stacked LSTM provides an output for each timestamp. Equations (8) to Equations (13) illustrate Nth layer of unrolled stacked LSTM.

$$\text{ForgetGate}: f_t^N = \sigma(W_f^N \cdot [h_{t-1}^N, h_t^{N-}] + b_f^N) \tag{8}$$

$$\text{InputGate}: i_t^N = \sigma(W_i^N \cdot [h_{t-1}^N, h_t^{N-1}] + b_i^N) \tag{9}$$

$$\text{ProcessInput:}: \hat{C}_t^N = \tanh\,(W_c^N \cdot [h_{t-1}^N, h_t^{N-1}] + b_c^N) \tag{10}$$

$$\text{CellUpdate}: C_t^N = f_t^N \star C_{t-1}^N + i_t^N \star \hat{C}_t^N \tag{11}$$

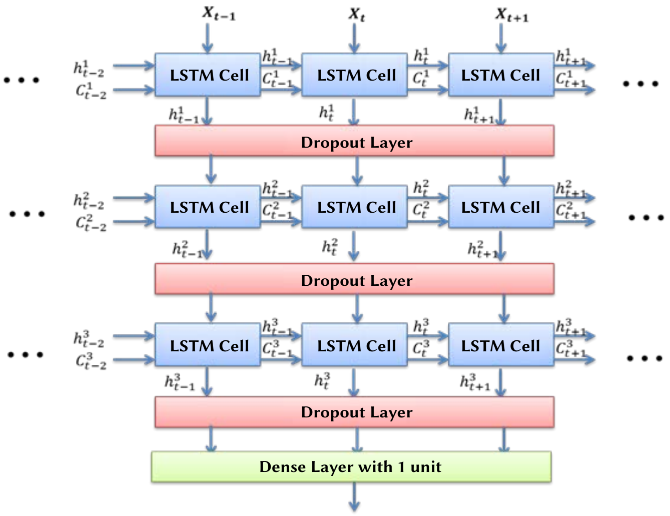


Fig. 5. Detail three-layer LSTM architecture.

$$\text{OutputGate}: o_t^N = \sigma(W_o^N \cdot [h_{t-1}^N, h_t^{N-1}] + b_o^N) \tag{12}$$

$$\text{Output}: h_t^N = o_t^N \star \tanh\,(C_t^N) \tag{13}$$

A stacked LSTM neural network outperform a regular LSTM due to its increased capacity and ability to capture more complex patterns in the data. Here are some reasons why a three-layer LSTM could outperform a regular LSTM:

- **Hierarchical Feature Extraction**: A stacked LSTM network has an additional layer to extract hierarchical features from the data. This allows it to learn abstract representations of the input data at different levels of granularity. The first layer can capture low-level features, the second layer can capture mid-level features, and the third layer can capture high-level features. This hierarchical feature extraction can make the network more effective at capturing complex patterns.
- **Increased Capacity**: The additional layer in a stacked LSTM increases the network's capacity to learn and represent data. It can model more intricate relationships and dependencies within the data, which is particularly useful for tasks involving long sequences or complex temporal dependencies.
- **Improved Generalization**: A deeper network, such as a three-layer LSTM, can sometimes generalize better to unseen data. While a deeper network has the potential to overfit the training data, appropriate regularization techniques (e.g., dropout) can help mitigate this risk. With proper regularization, a three-layer LSTM can learn to generalize well to new, unseen examples.
- **Better Representation Learning**: The stacked architecture allows for more sophisticated representation learning. Each layer can transform the input data into a more meaningful representation. This can be especially advantageous for tasks where the input data is complex or contains multiple levels of abstraction.

B. Cross-Validation

In this experiment we used rolling cross-validation method. It begin with a small initial subset of the data for training. Then, make forecasts for the subsequent data points and assess the accuracy of these forecasts. Importantly, the data points that were forecasted are subsequently incorporated into the training dataset for the next iteration. This process is repeated, allowing for the inclusion of previously forecasted data points in each subsequent training dataset, while forecasting the remaining data points. Rolling cross-validation provides a robust means of evaluating the performance of time-series models by continuously expanding the training dataset while testing against new, unseen data points. The experiment is performed on 5-fold-cross-validation as shown in Fig. 6.

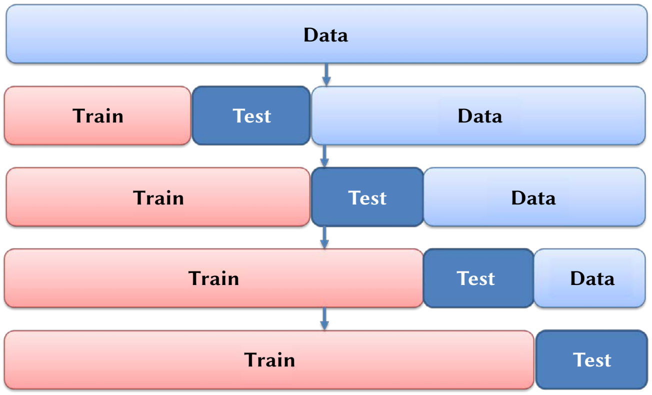


Fig. 6. Rolling cross-validation.

C. Evaluation Criteria

MAE and RMSE are selected for comprehensive and quantitative evaluation of prediction performance. RMSE and MAE represent absolute time series error and the scale is same as the data. A decrease in MAE and RMSE values signifies a reduction in model error residuals, indicating a higher level of accuracy in the prediction model. MAE and RMSE are expressed by Equation (14) and Equation (15) as,

$$\text{MAE} \;= \frac{1}{n}\sum_{t=1}^{n} |\hat{Y}_t - y_t| \tag{14}$$

$$RMSE \;= \sqrt{\frac{1}{n}\sum_{t=1}^{n} (\hat{Y}_t - y_t)^2} \tag{15}$$

where n is the total number of the predicted values $\hat{Y}_t$ and y_t is the actual value.

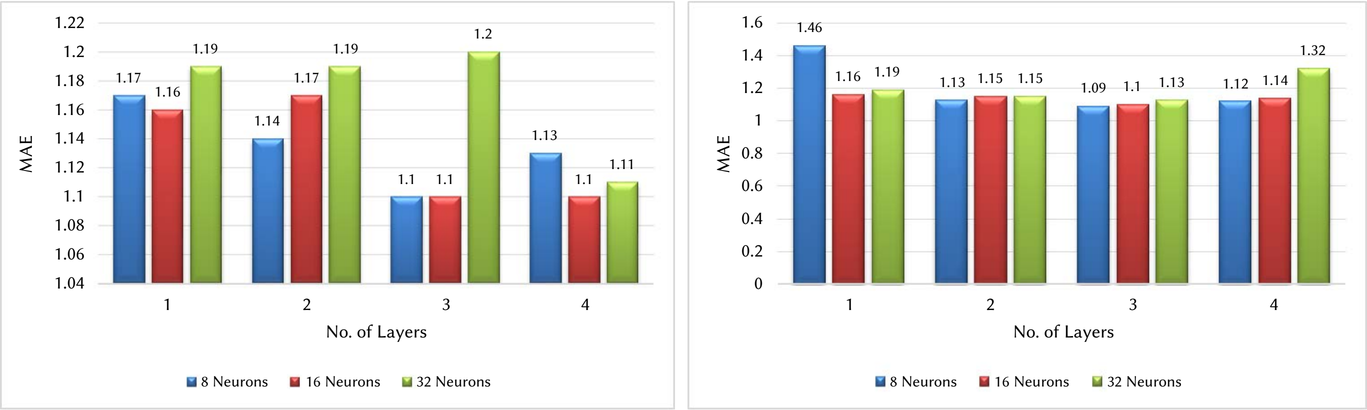
V. Experimental Analysis

In this section, results of proposed stacked LSTM on Sotavento and Yalova wind farm dataset are discussed. The proposed model is implemented using Python 3.10 employing Keras library. The loss function for LSTM was defined as the mean squared error, and the optimization method used was 'ADAM'. The dataset is partitioned into training and testing phases. Here the presented model follows training and testing ratio of 70:30. The performance of the forecasting models is assessed using RMSE and MAE.

A. Sotavento Wind Farm

In order to construct a stacked LSTM, trial and error method is used to select the size of the input layer and output layers. For that, experiments have been conducted on LSTM networks with 1 to 4 hidden layers, and the best LSTM network with the smallest MAE and RMSE has been selected.

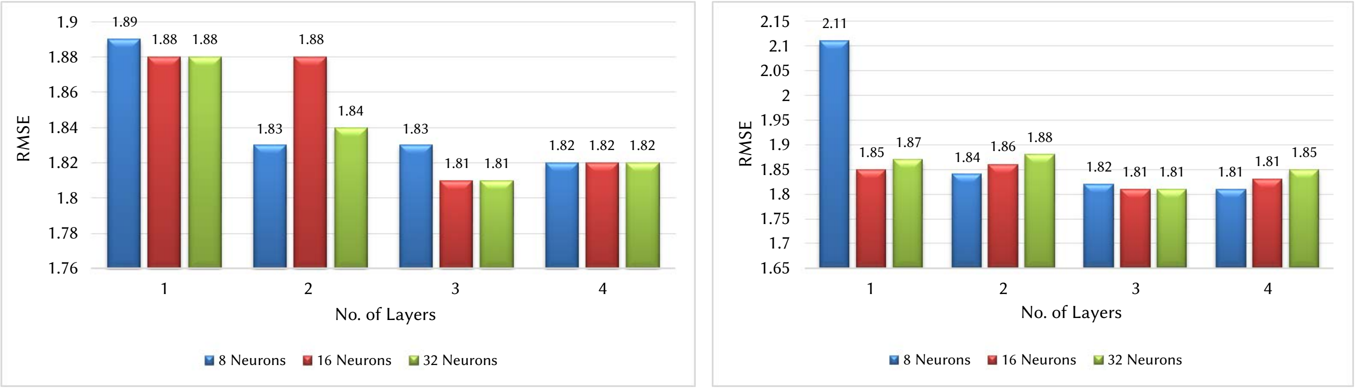
The model with the structure "16-16-16-1" works best. The proposed stacked LSTM shows a comparatively lesser MAE and RMSE values as depicted in Table III. When fourth layer is added, MAE and RMSE value start increasing because of number of parameters. It is also observed that when the missing values are imputed with the help of RFR, the performance is improved in every case as shown in Fig. 7b and Fig. 8b. Only for single layer it has slightly increases.



(a) MAE value after removing NA

(b) MAE value after imputing missing value using RFR

Fig. 7. MAE value of Sotavento wind farm dataset.



(a) RMSE value after removing NA

(b) RMSE value after imputing missing value using RFR

Fig. 8. RMSE value of Sotavento wind farm dataset.

TABLE III. MAE and RMSE Value of Sotavento Wind Farm

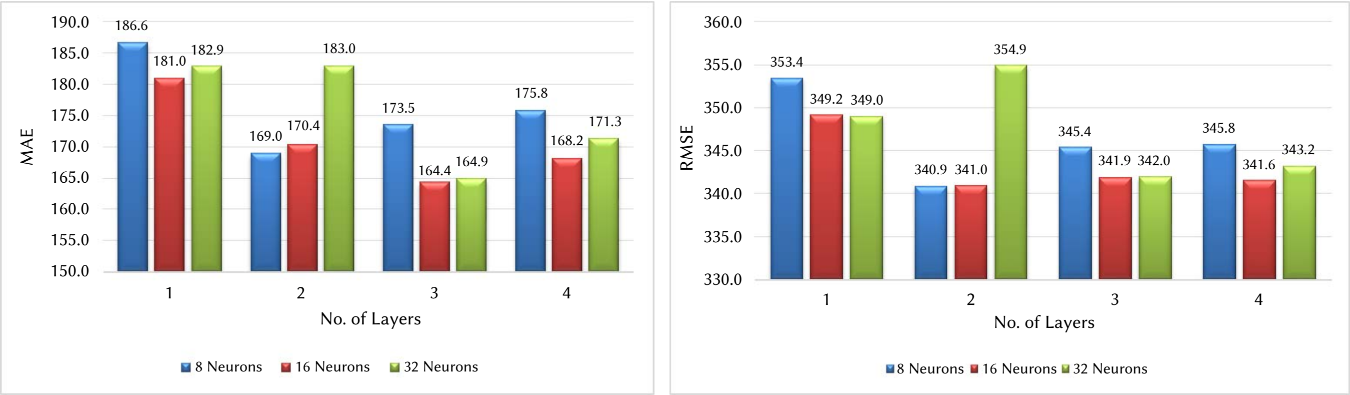
No. of Layers	No. of Neurons	Removing NA records		Missing value imputation using RFR	
		MAE	RMSE	MAE	RMSE
1	8	1.17	1.89	1.46	2.11
2		1.14	1.83	1.13	1.84
3		1.1	1.83	1.09	1.82
4		1.13	1.82	1.12	1.81
1	16	1.16	1.88	1.1	1.85
2		1.17	1.88	1.15	1.86
3		1.1	1.81	1.1	1.81
4		1.1	1.82	1.14	1.83
1	32	1.19	1.88	1.19	1.87
2		1.19	1.84	1.15	1.88
3		1.2	1.81	1.13	1.81
4		1.11	1.82	1.32	1.85

B. Yalova Wind Farm

Yalova Wind Farm dataset does not contain any missing value or outliers. The dataset is a clean dataset so it is directly divided into train and test set. From the experimental results shown in Table IV it is observed that the prediction performance decreases along with increase of amount of layers. The MAE and RMSE values for each model are plotted in Fig. 9. The graph shows that architecture with 3 layers and 16 neurons provides the smallest error value compared to other architectures. The minimum MAE of this architecture is 164.35, and the RMSE is 341.92. Therefore, the structure of "16-16-16-1" is selected as the optimal model.

TABLE IV. MAE and RMSE Value of Sotavento Wind Farm

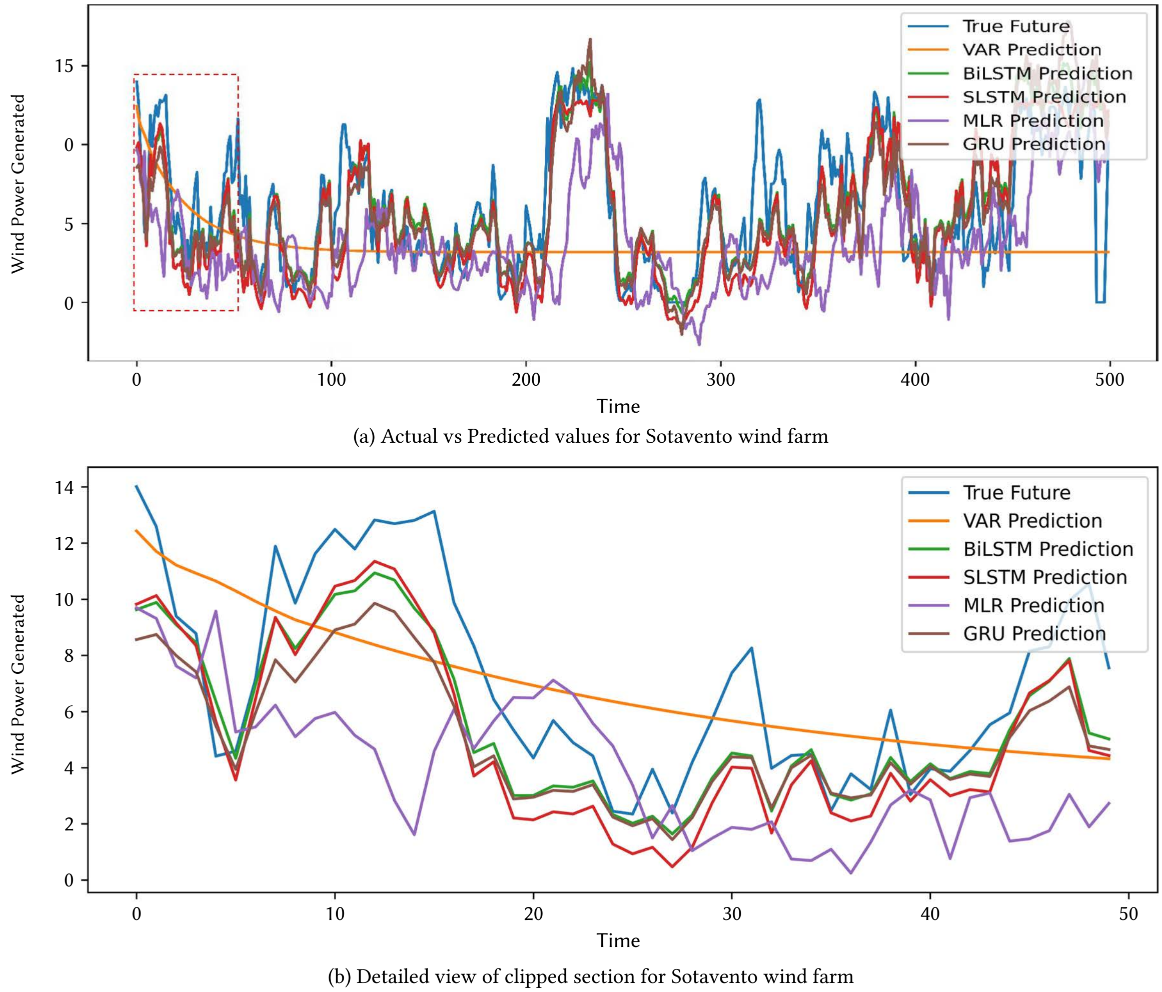
No. of Layers	No. of Neurons	Removing NA records	
		MAE	RMSE
1	8	186.63	353.44
2		168.97	340.87
3		173.54	345
4		175.82	345.77
1	16	180.96	349.15
2		170.38	340.99
3		164.35	341.92
4		168.15	341.57
1	32	182.88	349
2		182.98	354.92
3		164.89	341.97
4		171.34	343.21



(a) MAE value

(b) RMSE value

Fig. 9. MAE and RMSE value of YalovaWind Farm dataset.



(a) Actual vs Predicted values for Sotavento wind farm

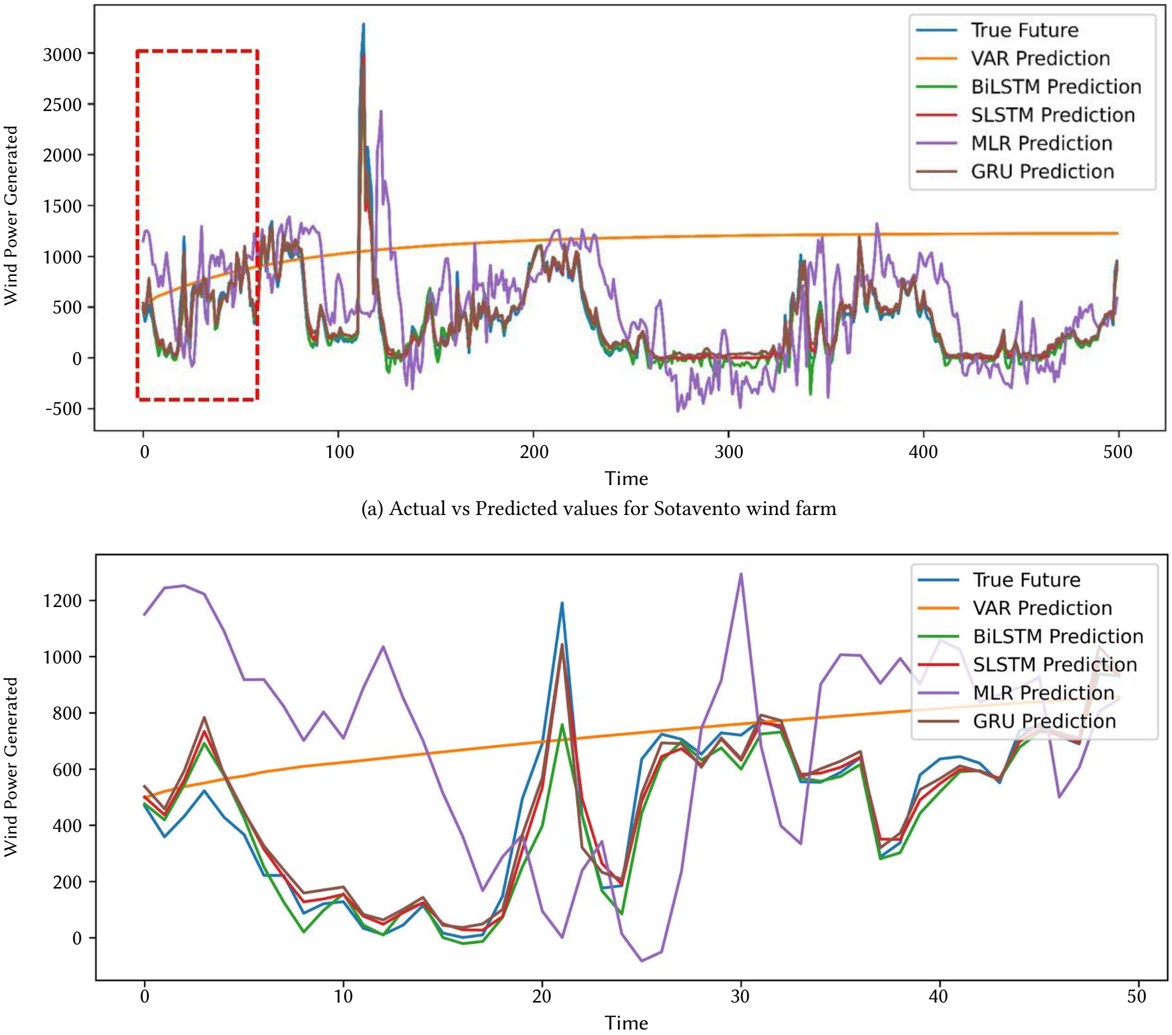
(b) Detailed view of clipped section for Sotavento wind farm

Fig. 10. Comparison of experimental results on Sotavento wind farm dataset.

C. Comparative Analysis

To verify the proposed models accuracy, four models including Multiple linear regression, VAR, GRU, BiLSTM are employed to conduct WPF of two wind farms. Table V shows the comparison results between proposed models and benchmark models. For regression model, the input is a multivariate time series of length ten (corresponding to 10 hours) and output is a real number corresponding to the next hour into the future. An input and hidden layers contain 16 neurons. Where as an output layer contains one neuron. The total number of epochs is 100, batch size is 64, and training rate is 0.001.

On the Yalova wind farm test dataset, the VAR model produces significantly higher MAE (995.09 and 1006.24) compared to the 3-layer stacked LSTM. This difference can be attributed to the VAR model's limitation in handling stationary time series data effectively. When we utilize the GRU model, the MAE increases by 0.07 and 25.04, and the RMSE increases by 0.05 and 7.28 for the Sotavento wind farm and Yalova Wind Farm, respectively, in comparison to the three-layer Stacked LSTM. The stacked LSTM model has much better performance than the baseline model, since it can capture longer temporal dependencies. Fig. 10a and Fig. 11a present graphical representations illustrating the



(a) Actual vs Predicted values for Sotavento wind farm

(b) Detailed view of clipped section for Sotavento wind farm

Fig. 11. Comparison of experimental results on Sotavento wind farm dataset.

predictions achieved by the five regressors for Sotavento Wind Farm and Yalova Wind Farm, respectively. The test result shows that, the proposed method is more robust and effective than the other five WPF methods. The fundamental reason for this is because the proposed method is based on the LSTM, which, due to the presence of self-feedback connections, is suited and successful in modelling time-series data. Subsequently, Fig. 10b and Fig. 11b provide a detailed view of the experiment. The detailed view shows the clarity of the improvements in measurements.

TABLE V. Comparative Analysis of RFR-Based Stacked LSTM

No.of Layers	Imputing Missing values using RFR		Yalova Wind Farm	
	MAE	RMSE	MAE	RMSE
Multiple Linear Regression	1.36	1.98	404.56	521.71
VAR	2.9	3.61	1159.44	1348.16
GRU	1.17	1.86	189.39	349.2
BiLSTM	1.75	2.28	180.44	342.94
Stacked LSTM	1.1	1.81	164.35	341.92

VI. Conclusion

Wind power is an essential area in renewable energy generation. This paper talks about the WPF utilizing a time series data acquired from wind mill. Here, the wind power is predicted effectively by employing the proposed stacked LSTM model. At first, utilizing two approaches such as removing missing values and RFR the input information is pre-processed. It is observed that the accuracy of the proposed model significantly increases when RFR is used for data imputation. It is followed by a 3-layer LSTM model referred as stacked LSTM to perform the task of WPF. From experimentation, it is observed that the proposed model achieved an average accuracy of 94.01% which is greater than the state-of-the-art approaches. The performance of the proposed model is evaluated using two error metrics such as MAE and RMSE. In addition, it is also observed that addition of a fourth layer slightly decreases the accuracy due to increased complexity. It is discovered from the analysis that the proposed model consistently outperformed other methods in terms of error. The proposed stacked LSTM model which is built on conventional LSTM, effectively captures the complex hidden patterns and changes in wind power output with respect to time due to its memory units and recurrent design. However, time required for training the data is comparatively high. Also, the performance of the proposed stacked LSTM requires large amount of data. In addition, finding an optimal hyperparameters is challenging and will require extensive experimentation.

This model can further be improved using a deep BiLSTM network. BiLSTMs are computationally more intensive than unidirectional LSTMs because they process the data in both directions. Furthermore, pre-processing methods which depict time-frequency analysis can be explored. Further research into the model's scalability, transferability to other contexts, and incorporation of additional relevant improvements could help in maximizing its practical and real time applications.

Acknowledgment

The authors wish to extend their appreciation to the Faculty Development Centre (VJTI-DBATU) focused on Geoinformatics, Spatial Computing, and Big Data Analytics. This center, established under the auspices of PMMMNMTT, MHRD, Government of India, New Delhi, has been involved in enhancing capacity in various fields, including Geoinformatics, AI & Machine Learning, Deep Learning, Big Data Analytics, and related domains. Their support has been invaluable in advancing our research in this study.

References

[1] S. K. Kim and S. Park, "Impacts of renewable energy on climate vulnerability: A global perspective for energy transition in a climate adaptation framework," *Science of The Total Environment*, vol. 859, pp. 160175, Feb. 2023, doi: 10.1016/J.SCITOTENV.2022.160175.

[2] S. M. H. D. Perera, G. Putrus, M. Conlon, M. Narayana, and K. Sunderland, "Wind Energy Harvesting and Conversion Systems: A Technical Review," *Energies*, vol. 15, no. 24, p. 1-34, 2022.

[3] T. M. Giannaros, D. Melas, I. Ziomas, "Performance evaluation of the Weather Research and Forecasting (WRF) model for assessing wind resource in Greece," *Renewable Energy*, vol. 102, pp. 190–198, 2017.

[4] C. Liu, X. Zhang, S. Mei, Z. Zhen, M. Jia, Z. Li, H. Tang, "Numerical weather prediction enhanced wind power forecasting: Rank ensemble and probabilistic fluctuation awareness," *Applied Energy*, vol. 313, p. 118769, 2022.

[5] S. Kumari, S. Sreekumar, S. Singh, D. P. Kothari, "Comparison Among ARIMA, ANN, and SVR Models for Wind Power Deviation Charge Reduction," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, vol. 1, 2022, pp. 551–557, IEEE.

[6] F. Yao, W. Liu, X. Zhao, L. Song, "Integrated Machine Learning and Enhanced Statistical Approach-Based Wind Power Forecasting in Australian Tasmania Wind Farm," *Complexity*, vol. 2020, 2020, doi: 10.1155/2020/9250937.

[7] Z. Wu, G. Luo, Z. Yang, Y. Guo, K. Li, Y. Xue, "A comprehensive review on deep learning approaches in wind forecasting applications," *CAAI Transactions on Intelligence Technology*, vol. 7, no. 2, pp. 129–143, 2022.

[8] Y. Wang, R. Zou, F. Liu, L. Zhang, Q. Liu, "A review of wind speed and wind power forecasting with deep neural networks," *Applied Energy*, vol. 304, p. 117766, 2021.

[9] H. Zhen, D. Niu, M. Yu, K. Wang, Y. Liang, and X. Xu, "A Hybrid Deep Learning Model and Comparison for Wind Power Forecasting Considering Temporal-Spatial Feature Extraction," *Sustainability*, vol. 12, no. 22, pp. 9490, May 2020, doi: 10.3390/SU12229490.

[10] K.-S. Chen, K.-P. Lin, J.-X. Yan, and W.-L. Hsieh, "Renewable Power Output Forecasting Using Least-Squares Support Vector Regression and Google Data," Sustainability, vol. 11, no. 11, pp. *3009*, May 2019, doi: 10.3390/SU11113009.

[11] J. Zhou, X. Yu, and B. Jin, "Short-Term Wind Power Forecasting: A New Hybrid Model Combined Extreme-Point Symmetric Mode Decomposition, Extreme Learning Machine and Particle Swarm Optimization," *Sustainability*, vol. 10, no.9, pp. 3202, May 2018, doi: 10.3390/SU10093202.

[12] B. Xiong, L. Lou, X. Meng, X. Wang, H. Ma, Z. Wang, "Short-term wind power forecasting based on Attention Mechanism and Deep Learning," *Electric Power Systems Research*, vol. 206, 2022.

[13] L. Ye, B. Dai, M. Pei, P. Lu, J. Zhao, M. Chen, B. Wang, "Combined approach for short-term wind power forecasting based on wave division and Seq2Seq model using deep learning," *IEEE Transactions on Industry Applications*, vol. 58, no. 2, pp. 2586–2596, 2022.

[14] A. Alkesaiberi, F. Harrou, Y. Sun, "Efficient wind power prediction using machine learning methods: A comparative study," *Energies*, vol. 15, no. 7, p. 2327, 2022.

[15] J. Jamii, M. Mansouri, M. Trabelsi, M. F. Mimouni, W. Shatanawi, "Effective artificial neural network- based wind power generation and load demand forecasting for optimum energy management," *Frontiers in Energy Research*, vol. 10, 2022.

[16] D.-D. Yuan, M. Li, H.-Y. Li, C.-J. Lin, B.-X. Ji, "Wind power prediction method: Support vector regression optimized by improved jellyfish search algorithm," *Energies*, vol. 15, no. 17, p. 6404, 2022.

[17] X. Shi, X. Lei, Q. Huang, S. Huang, K. Ren, Y. Hu, "Hourly Day-Ahead Wind Power Prediction Using the Hybrid Model of Variational Model Decomposition and Long Short-Term Memory," *Energies 2018, Vol. 11, Page 3227*, vol. 11, p. 3227, may 2018, doi: 10.3390/EN11113227.

[18] K. Qu, G. Si, Z. Shan, X. Kong, X. Yang, "Short- term forecasting for multiple wind farms based on transformer model," *Energy Reports*, vol. 8, pp. 483–490, 2022.

[19] E. G. S. Nascimento, T. A. de Melo, D. M. Moreira, "A transformer-based deep neural network with wavelet transform for forecasting wind speed and wind energy," *Energy*, vol. 278, pp. 127–678, 2023.

[20] A. Lagos, J. E. Caicedo, G. Coria, A. R. Quete, M. Martz'zinez, G. Suvire, J. Riquelme, "State-of-the- Art using bibliometric analysis of Wind-Speed and- Power forecasting methods applied in power systems," *Energies*, vol. 15, no. 18, p. 6545, 2022.

[21] A. Tascikaraoglu, M. Uzunoglu, "A review of combined approaches for prediction of short-term wind speed and power," 2014, doi: 10.1016/j.rser.2014.03.033.

[22] E. López, C. Valle, H. Allende, E. Gil, H. Madsen, "Wind Power Forecasting Based on Echo State Networks and Long Short-Term Memory," *Energies 2018, Vol. 11, Page 526*, vol. 11, p. 526, may 2018, doi: 10.3390/EN11030526.

[23] F. Shahid, A. Zameer, A. Mehmood, M. A. Z. Raja, "A novel wavenets long short term memory paradigm for wind power prediction," *Applied Energy*, vol. 269, pp. 115098, may 2020, doi: 10.1016/J.APENERGY.2020.115098.

[24] J. F. Toubeau, P. D. Dapoz, J. Bottieau, A. Wautier, Z. D. Grève, F. Vallée, "Recalibration of recurrent neural networks for short-term wind power forecasting," *Electric Power Systems Research*, vol. 190, pp. 106639, may 2021, doi: 10.1016/J.EPSR.2020.106639.

[25] Z. Sun, M. Zhao, "Short-Term Wind Power Forecasting Based on VMD Decomposition, ConvLSTM Networks and Error Analysis," *IEEE Access*, vol. 8, pp. 134422– 134434, 2020, doi: 10.1109/ACCESS.2020.3011060.

[26] S. Siami-Namini, N. Tavakoli, A. S. Namin, "A Comparison of ARIMA and LSTM in Forecasting Time Series," *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, pp. 1394–1401, may 2019, doi: 10.1109/ICMLA.2018.00227.

[27] H. S. Dhiman, P. Anand, D. Deb, "Wavelet transform and variants of SVR with application in wind forecasting," *Advances in Intelligent Systems and Computing*, vol. 757, pp. 501–511, 2019, doi: 10.1007/978-981-13-1966-245.

[28] Y. Zhang, J. Le, X. Liao, F. Zheng, Y. Li, "A novel combination forecasting model for wind power integrating least square support vector machine, deep belief network, singular spectrum analysis and locality- sensitive hashing," *Energy*, vol. 168, pp. 558–572, may 2019, doi: 10.1016/J.ENERGY.2018.11.128.

[29] S. Han, Y. Liu, J. Li, "Wind power combination prediction based on the maximum information entropy principle," 2012. [Online]. Available: https://ieeexplore.ieee.org/document/6321153.

[30] M. I. A. A. Khalaf, J. Q. Gan, "Deep classifier structures with autoencoder for higher-level feature extraction," *IJCCI 2018 - Proceedings of the 10th International Joint Conference on Computational Intelligence*, pp. 31–38, 2018, doi: 10.5220/0006883000310038.

[31] C. Fu, G.-Q. Li, K.-P. Lin, and H.-J. Zhang, "Short-Term Wind Power Prediction Based on Improved Chicken Algorithm Optimization Support Vector Machine," Sustainability, vol. 11, pp. 512, May 2019, doi: 10.3390/SU11020512.

[32] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine

translation," *arXiv preprint arXiv:1609.08144*, 2016.

[33] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, pp. 9, 2019.

[34] J. Shah, D. Vaidya, and M. Shah, "A comprehensive review on multiple hybrid deep learning approaches for stock prediction," *Intelligent Systems with Applications*, vol. 16, 2022, https://doi.org/10.1016/j.iswa.2022.200111.

[35] R. K. Behera, S. Das, S. K. Rath, S. Misra, R. Damasevicius, "Comparative study of real time machine learning models for stock prediction through streaming data.," *The Journal of Universal Computer Science*, vol. 26, no. 9, pp. 1128–1147, 2020.

[36] A. M. AlRassas, M. A. Al-qaness, A. A. Ewees, S. Ren, M. Abd Elaziz, R. Damaševičius, T. Krilavičius, "Optimized anfis model using aquila optimizer for oil production forecasting," *Processes*, vol. 9, no. 7, p. 1194, 2021.

[37] B. Y. El-Habil, S. S. Abu-Naser, "Global climate prediction using deep learning," *Journal of Theoretical and Applied Information Technology, vol. 100, no. 24*, 2022.

[38] J. Oruh, S. Viriri, A. Adegun, "Long short-term memory recurrent neural network for automatic speech recognition," *IEEE Access*, vol. 10, pp. 30069– 30079, 2022.

[39] D. Chen, J. Wen, C. Lv, "A spatio-temporal attention graph convolutional networks for sea surface temperature prediction," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, pp. 64–72, 2023, doi: https://doi.org/10.9781/ijimai.2023.02.011.

[40] M. Martínez-Comesaña, J. Martínez-Torres, P. Eguía- Oller, J. López-Gómez, "Use of optimised lstm neural networks pre-trained with synthetic data to estimate pv generation," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, pp. 1–10, 2023, doi: https://doi.org/10.9781/ijimai.2023.11.002.

[41] M. Roy, S. Majumder, A. Halder, U. Biswas, "Ecg-net: A deep lstm autoencoder for detecting anomalous ecg," *Engineering Applications of Artificial Intelligence*, vol. 124, 2023.

[42] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[43] F. J. Ordóñez, D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, pp. 115, 2016.

[44] H. N. Akouemo, R. J. Povinelli, "Data improving in time series using ARX and ANN models," *IEEE Transactions on Power Systems*, vol. 32, no. 5, pp. 3352– 3359, 2017.

[45] F. Harlé, F. Chatelain, C. Gouy-Pailler, S. Achard, "Bayesian model for multiple change-points detection in multivariate time series," *IEEE Transactions on Signal Processing*, vol. 64, no. 16, pp. 4351–4362, 2016.

[46] O. Alghushairy, R. Alsini, T. Soule, X. Ma, "A review of local outlier factor algorithms for outlier detection in big data streams," *Big Data and Cognitive Computing*, vol. 5, no. 1, 2020.

[47] M. Galphade, V. Nikam, B. Banerjee, A. Kiwelekar, "Intelligent multiperiod wind power forecast model using statistical and machine learning model," *Bulletin of Electrical Engineering and Informatics*, vol. 11, pp. 1186–1193, may 2022, doi: 10.11591/EEI.V11I3.3756.

[48] T. Carpenito, J. Manjourides, "MISL: Multiple imputation by super learning," *Statistical methods in medical research*, vol. 31, no. 10, pp. 1904–1915, 2022.

Manisha Galphade

Manisha Galphade is Bachelor, and Masters in Engineering (Computer Science and Engineering) pursuing PhD in Computer Department of VJTI. She has 14 years of teaching experience. Her research interest includes Machine Learning, Deep learning, GIS, Geospatial Analysis, Weather Predictions, Wind Power Prediction Modelling, Data mining.

Dr V.B.Nikam

Dr V. B. Nikam Associate Professor, Computer Engg & IT,VJTI Mumbai, has done Bachelors,Masters and PhD in Computer Engineering.He has 25 yrs experience, guided 50+ PG,25+ UG projects,and Supervised 4 PhDs. He was felicitated with IBM TGMC-2010 DRONA award by IBM Academic initiatives. He is Senior Member (CSI),Senior Member (IEEE), Senior Member (ACM). He worked onBARC ANUPAM Supercomputer. He was invited to JAPAN for a K-Supercomputer study tour in 2013. He has received a grant-in-aid from NVIDIA for CUDA Teaching and Research,2013.Presently,he is PI & Coordinator, Faculty Development Center (Geo-informatics, Spatial Computing and BigData Analytics) funded by MHRD, Govt of India. He works in the area of Data Mining and Data Warehousing, Machine Learning, Geoinformatics,Big Data Analytics, Geo-Spatial Analysis,Cloud Computing, GPUHigh Performance Computing. You may visit his webpage www.drvbnikam.in.

Dr. Biplab Banerjee

Dr. Biplab Banerjee received the M.E. (Computer Science and Engineering) from Jadhavpur University, Kolkata, and Ph.D. in Satellite Image Analysis from the Indian Institute of Technology Bombay, Mumbai, India. Dr. Banerjee received the Excellence in Ph.D. Thesis Award for his Ph.D. thesis from the IIT Bombay. He is Postdoctoral Researcher at the University of Caen Basse-Normandy, France and the Istituto Italiano di Technologia Genova, Italy. He is engaged in many research projects including international collaborated projects. His interests include computer vision and machine learning. Dr Biplab currently supervising 9 PhD students, and has guided more than 25 MTech thesis so far. He is reviewer for IEEE journals, and Transactions in Image Processing, Applied Earth Observations and Remote Sensing, Neural Computing & Applications (Springer), Journal of Indian Society of Remote Sensing (Springer), Computer Vision and Image Understanding (Elsevier). He is Member of IEEE. Indo-Canadian Research Grants, Board of Research in Nuclear Sciences – Department of Atomic Energy, Vishveshwarya PhD Scheme, Institute for Plasma Research, GUJCOST-DST and many other MNCs.

Dr. Arvind W. Kiwelekar

Dr. Arvind W. Kiwelekar Professor in Computer Science, Dr. B. A. Technological University Lonere has done Ph. D. from Indian Institute of Technology, Mumbai. He has 26 years of experience. His research interest includes Software Engineering, Software Architecture, Applied Artificial Intelligence and Ontology.

Dr. Priyanka Sharma

Dr Priyanka Sharma has 24 years of experience that spans both Industry and Academia, she specialize in leading AI and Data Analytics based application development and integrated solution design for various inter-disciplinary domains. She is a NVIDIA DLI Ambassador and have trained more than 3000 professionals through NVIDIA Led hands-on training programs on Deep Learning, Computer Vision, Natural Language Processing and CUDA Programming. She has also been associated with other international firms as Corporate Trainer apart from being a passionate academician and professor in CSE Department, Nirma University. She has published more than 55 research papers in International Journal, Books and Conferences in the domain of Artificial Intelligence and Deep Learning. She was earlier Principal Investigator/Collaborator of NVIDIA Research Center at Nirma University and several other research projects funded under Shastri

Use of Data Mining for Intelligent Evaluation of Imputation Methods

David L. la Red Martínez[1]*, Carlos R. Primorac[2]

[1] National Technological University, Resistencia Regional Faculty, Resistencia (Argentina)
[2] Computer Science Department, National University of the Northeast, Corrientes (Argentina)

* Corresponding author: lrmdavid@ca.frre.utn.edu.ar

Received 6 February 2022 | Accepted 1 March 2023 | Published 23 March 2023

Abstract

In real-world situations, researchers frequently face the difficulty of missing values (MV), i.e., values not observed in a data set. Data imputation techniques allow the estimation of MV using different algorithms, by means of which important data can be imputed for a particular instance. Most of the literature in this field deals with different imputation methods. However, few studies deal with a comparative evaluation of the different methods as to provide more appropriate guidelines for the selection of the method to be applied to impute data for specific situations. The objective of this work is to show a methodology for evaluating the performance of imputation methods by means of new metrics derived from data mining processes, using quality metrics of data mining models. We started from the complete dataset that was amputated with different amputation mechanisms to generate 63 datasets with MV; these were imputed using Median, k-NN, k-Means and Hot-Deck imputation methods. The performance of the imputation methods was evaluated using new metrics derived from quality metrics of the data mining processes, performed with the original full file and with the imputed files. This evaluation is not based on measuring the error when imputing (usual operation), but on considering the similarity of the values of the quality metrics of the data mining processes obtained with the original file and with the imputed files. The results show that –globally considered and according to the new proposed metric, the imputation methods that showed the best performance were k-NN and k-Means. An additional advantage of the proposed methodology is that it provides predictive data mining models that can be used *a posteriori.*

Keywords

Computer Science, Data Imputation, Data Mining, Interdisciplinary Applications, Performance Evaluation of Imputation Methods.

DOI: 10.9781/ijimai.2023.03.002

I. Introduction

MVS (Missing Values) introduce an element of ambiguity in data analysis. They can affect the properties of statistical estimators such as mean, variance or percentages, resulting in a loss of power and false conclusions. Data imputation is an alternative to deal with MV. Most of the published work in this field deals with the development of new imputation methods. However, few studies report a comprehensive evaluation of existing methods to provide guidelines to make the most appropriate methodological choice in practice [1].

The literature proposes two general approaches to dealing with MVs [2]. In the simplest case, they are omitted. A second option is to use imputation techniques and, from the complete data, estimate them using different algorithms, whereby an important feature can be imputed for a particular instance [3].

A classical approach to performance evaluation of imputation methods is described in [4].

Other works have proposed the use of machine learning (ML) algorithms as imputation methods [5]. These techniques are based on building a predictive model to estimate missing data based on the available values in the dataset [6]. In [5], the suitability of supervised (classification) and unsupervised (clustering) learning algorithms for imputation is studied. ML algorithms such as decision trees (DT), k-Nearest Neighbors (k-NN), k-Means Clustering and Bayesian Networks have been used as imputation methods in different domains [5], [6], [7], [8], [9], [10].

In this work, a continuation of [11], we do not propose the use of ML and data mining (DM) algorithms to impute. We rather propose an innovative criterion to evaluate the performance of imputation methods (IM), in this case Medians, k-NN, k-Means and Hot-Deck, using the value of quality indicator metrics of a data mining model (DMM) obtained through data mining processes (DMP). The polynomial regression technique was used to create predictive DMMs.

Please cite this article in press as:
D. L. la Red Martínez, C. R. Primorac. Use of Data Mining for Intelligent Evaluation of Imputation Methods, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 82-95, 2025, http://dx.doi.org/10.9781/ijimai.2023.03.002

We use the criterion of highest similarity between the results of the data mining processes using the original dataset (with complete data) and the imputed datasets after being amputated. New specific metrics were defined from the values of the metrics obtained by the DMPs.

We used the original "Iris" data set and 252 data sets imputed after amputation.

Quality, accuracy (precision) and classification metrics were considered as indicators of DMM quality [12].

The article is organized as follows: the Data Mining (DM) concept review section introduces the main algorithms and model evaluation metrics, the Materials and methods section describes the datasets, the data mining algorithm and the quality indicator metrics used, the Results and discussions section discusses and compares them in detail, and concludes with Conclusions, Future work, Acknowledgements and References.

II. Review of Data Mining Concepts (DM)

Historically, the notion of discovering hidden patterns in data has been given a variety of names including data mining (DM) and knowledge discovery (KDD: Knowledge Discovery in Databases). KDD refers to the general process of discovering useful knowledge from data. KDD is the application of specific algorithms to extract patterns from data. DM is a stage within the general KDD process that refers to the algorithmic means by which patterns are extracted and enumerated from data [13].

The generation of a DMM is part of a larger process that includes from the formulation of questions about the data and the creation of a model to answer them, to the implementation of the model in a working environment. In a broad sense, DMP can be defined by the following basic steps: data acquisition, preprocessing, model generation, evaluation, and exploitation [14].

In addition, DMP is cyclical in nature, meaning that the generation of a DMM is a dynamic and iterative process [15], [16].

A. Generation of DM Models

In practice, the two main objectives of DM, prediction and description, can be achieved by using a variety of methods [17].

Predictive methods include supervised learning techniques such as classification and regression. Descriptive methods include unsupervised learning techniques such as clustering, association rules or sequence discovery [12].

A DM algorithm is a set of calculations and heuristic rules that allows the creation of a DMM from data. To create a model, the algorithm first analyzes the data provided, looking for specific types of patterns or trends. The algorithm uses the results of this analysis to define the optimal parameters for creating the DMM. These parameters are then applied across the entire dataset to extract actionable patterns and detailed statistics [14].

The most common classification techniques include tree algorithms and decision rules, Bayesian classifiers, nearest neighbor-based classifiers, logistic regression, support vector machines (SVM) and artificial neural networks (ANN) [12], [15], [18].

The most common regression techniques include linear regression algorithms (simple and multiple), polynomial and weighted local regression, regression trees, SVM for regression and ANN [19], [12], [18].

In general, the main clustering algorithms include partitioning, hierarchical, distance-based and mesh-based methods [15].

The performance evaluation of a DMM is probably the most critical step in the entire DMP [16].

The quality of classification models is often assessed by the classification accuracy and the confusion matrix [18].

In regression problems, measures are based on the difference between the true value and the value predicted by the model [18].

III. Materials and Methods

This section describes the procedure followed to evaluate the performance of four imputation methods (IM): Medians, k-NN, k-Means and Hot-Deck, using the values of quality, accuracy (precision) and classification metrics obtained through data mining processes, using polynomial regression models to classify the "Iris" plant type.

A. Data Mining

IBM InfoSphere Warehouse (ISW) V.9.7 software was used, which includes, among others, tools (Intelligent Miner, Design Studio, etc.), for the creation, interpretation, and evaluation of DMM [20].

The original "Iris" data set and 252 imputed "Iris" data sets, obtained from imputing by Mean, k-NN, k-Means and Hot-Deck IM the amputated data sets in the 63 combinations of mechanisms, patterns and MV percentages, as thoroughly detailed in [11], were used.

In the DM stage, the techniques to be used were selected and the corresponding mining flows were created, in which the respective algorithms were parameterized.

The polynomial regression technique was considered. Its objective is to predict the numerical value of the dependent variable on known values thus creating models that can then be used to predict new or unknown values.

For the analysis of results, the "Iris" data set was considered, corresponding to the plant species of the same name. The type of plant was selected as the objective variable *t* and the width and length of petals and sepals as independent variables, as presented in Table I.

TABLE I. "Iris" Correlation Matrix [11]

	sepal. length	**sepal. width**	**petal. length**	**petal. width**	**class**
sepal. length	1.0000	-0.1777	0.8774	0.8288	0.7885
sepal. width	-0.1777	1.0000	-0.4434	-0.3549	-0.4320
petal. length	0.8774	-0.4434	1.0000	0.9619	0.9462
petal. width	0.8288	-0.3549	0.9619	1.0000	0.9526
class	0.7885	-0.4320	0.9462	0.9526	1.0000

In Design Studio, DMPs are performed by creating and executing DM flows. The design of a flow includes, at a minimum, an input table operator, and a DM operator specific to the DM technique being used. Additionally, most flows include one or more output operators, such as the visualization operator that presents the value of the metrics to evaluate the obtained model [20].

The DM flow used to perform the DMP has the following structure: The <Source Table> operator defines the data set, which in this case consists of one record for each sample of the "Iris" plant file, composed of the four predictor attributes and the target attribute described in Table I. The <Predictor> operator executes the indicated DM algorithm (polynomial regression) and sends the obtained DMM to the <Visualizer> operator, which finally presents the information to evaluate the DMP result.

The model quality metrics, which range from 0 to 1, are presented by the Design Studio viewer operator, and considered to evaluate the quality of the DMM obtained in each of the DMPs: i) model *quality*, ii) *accuracy (precision)* and iii) *classification* [12].

Model quality compares the model's predictive performance with the predictive performance of a trivial model that always returns the mean of the target attribute as the prediction value. A quality value of zero indicates that the model's predictive performance is no better than predicting the standard. In contrast, a value close to one indicates that the model's predictive performance is far superior to predicting the mean [12].

Accuracy (precision) represents the probability that a prediction be correct [12].

Finally, the *classification* is a measure of the model's ability to sort the records correctly. It is calculated according to the order of the test set records when sorted by the predicted values with the order of the same data records when sorted by the actual values of the target variable [12].

DM flows were run for the original "Iris" dataset and the 252 datasets imputed by the Means, k-NN, k-Means and Hot-Deck IM, after being amputated in the different amputation combinations described in [11].

For each DM flow, the values of three metrics indicating the quality of the DMM achieved were obtained.

B. Evaluation of the Performance of Imputation Methods (IM) Using Metrics Obtained From Data Mining Processes (DMP)

It is considered:

- The data set Y shown in Table II, with n cases and p variables, where y_{ij} are observed values, with $1 \leq i \leq n$ and $1 \leq j \leq p$.
- The imputed data sets $Y^{a_r m_s}$ depicted in Table III, with $1 \leq r \leq l$ and $1 \leq s \leq t$.
- The metrics q_i indicated in Table IV, which are quality indicators of DMM, with $1 \leq i \leq k$.

Table IV shows the values of the metrics q_i, with $1 \leq i \leq k$, which are the quality indicators of the DMM obtained by the DMP using the data set Y.

TABLE II. Original Data Set Y [11]

Y_1	Y_2	...	Y_j	...	Y_p
y_{11}	y_{12}	...	y_{1j}	...	y_{1p}
y_{21}	y_{22}	...	y_{2j}	...	y_{2p}
...	...	...	...	...	...
y_{i1}	y_{i2}	...	y_{ij}	...	y_{ip}
...	...	...	...	...	...
y_{n1}	y_{n2}	...	y_{nj}	...	y_{np}

TABLE III. Datasets Y^{ARMS} With Elements $y_{ij}^{a_r m_s}$ Imputed by the M_s Method After Having Been Amputated by the A_R Mechanism [11]

$Y_1^{a_r m_s}$	$Y_2^{a_r m_s}$	...	$Y_j^{a_r m_s}$	...	$Y_p^{a_r m_s}$
$y_{11}^{a_r m_s}$	$y_{12}^{a_r m_s}$	...	$y_{1j}^{a_r m_s}$	...	$y_{1p}^{a_r m_s}$
$y_{21}^{a_r m_s}$	$y_{22}^{a_r m_s}$	...	$y_{2j}^{a_r m_s}$	...	$y_{2p}^{a_r m_s}$
⋮	⋮	...	⋮	...	⋮
$y_{i1}^{a_r m_s}$	$y_{i2}^{a_r m_s}$	...	$y_{ij}^{a_r m_s}$	...	$y_{ip}^{a_r m_s}$
⋮	⋮	...	⋮	...	⋮
$y_{n1}^{a_r m_s}$	$y_{n2}^{a_r m_s}$	...	$y_{nj}^{a_r m_s}$	...	$y_{np}^{a_r m_s}$

TABLE IV. Values of the Metrics $q_i(Y)$ Indicating the Quality of the DMM (Own Elaboration)

	q_1	q_2	...	q_i	...	q_k
Y	$q_1(Y)$	$q_2(Y)$	...	$q_i(Y)$	...	$q_k(Y)$

It is considered $q_i(Y^{a_r m_s})$ the values of the q_i metrics, indicators of quality of the DMM obtained through the DMP using the data sets $Y^{a_r m_s}$, with $1 \leq r \leq l$ and $1 \leq s \leq t$, represented in Table V.

The metric Δq_i^{rs}, with $1 \leq i \leq k$; $1 \leq r \leq l$ and $1 \leq s \leq t$, equation (1), was defined. That is, the difference in absolute value, between the values of the metrics $q_i(Y)$ and $q_i(Y^{a_r m_s})$ represented in Tables IV and V respectively.

Thus, with respect to the Δq_i^{rs} metric, the best imputation method m_s, with $1 \leq s \leq t$, for imputing the amputated Y data set in the combination a_r, with $1 \leq r \leq l$, is the one that minimizes the value of the Δq_i^{rs} metric, with $1 \leq i \leq k$.

TABLE V. Values of $q_i(Y^{a_r m_s})$ (Own Elaboration)

Y	m_1	...	m_1	...	m_s	...	m_s	...
a_1	$q_1(Y^{a_1 m_1})$	...	$q_k(Y^{a_1 m_1})$	...	$q_1(Y^{a_1 m_s})$	...	$q_k(Y^{a_1 m_s})$	...
a_2	$q_1(Y^{a_2 m_1})$	...	$q_k(Y^{a_2 m_1})$	...	$q_1(Y^{a_2 m_s})$	...	$q_k(Y^{a_2 m_s})$	...
...	...	...	...	...	...	...	...	...
a_r	$q_1(Y^{a_r m_1})$	...	$q_k(Y^{a_r m_1})$	...	$q_1(Y^{a_r m_s})$	...	$q_k(Y^{a_r m_s})$	...
...	...	...	...	...	...	...	...	...
a_l	$q_1(Y^{a_l m_1})$	...	$q_k(Y^{a_l m_1})$	...	$q_1(Y^{a_l m_s})$	...	$q_k(Y^{a_l m_s})$	...

Table VI summarizes the values as expressed in equation (1).

$$\Delta q_i^{rs} = |q_i(Y) - q_i(Y^{a_r m_s})| \tag{1}$$

TABLE VI. Δq_i^{rs} Values (Own Elaboration)

Y	m_1	...	m_1	...	m_s	...	m_s	...
a_1	Δq_1^{11}	...	Δq_k^{11}	...	Δq_1^{1s}	...	Δq_k^{1s}	...
a_2	Δq_1^{21}	...	Δq_k^{21}	...	Δq_1^{2s}	...	Δq_k^{2s}	...
...	...	...	...	...	...	...	...	...
a_r	Δq_1^{r1}	...	Δq_k^{r1}	...	Δq_1^{rs}	...	Δq_k^{rs}	...
...	...	...	...	...	...	...	...	...
a_l	Δq_1^{l1}	...	Δq_k^{l1}	...	Δq_1^{ls}	...	Δq_k^{ls}	...

Thus, by ascendingly ordering the imputation methods by the values given by equation (1), we obtain the order of goodness of fit of the m_s, with $1 \leq s \leq t$, imputation methods used to impute the amputated Y data set in the combination a_r, with $1 \leq r \leq l$, with respect to the metric Δq_i^{rs}, with $1 \leq i \leq k$.

The performance of the imputation methods used to impute an amputated data set was evaluated using this newly defined metric, which made it possible to obtain an order of goodness of imputation methods, considering an evaluation criterion.

The order of goodness of imputation methods with respect to the criterion considered was defined as an ordered list or ratio of imputation methods according to their performance in imputing an amputated data set, considering an evaluation criterion. In this list, the best method is ranked first and the worst last.

In this scenario, the best imputation method according to one criterion (and its corresponding metric) may turn out to be the worst according to the remaining criteria. Evaluating an imputation method according to a single metric may not be sufficient, as the best method in terms of two or more metrics simultaneously may be of interest.

An aggregation operator makes it possible to aggregate, merge or synthesize information, that is, to combine a series of data from different sources to reach a certain conclusion or make a certain decision [21], [22].

In order to find the best imputation method m_s to impute an amputated *data set* in the combination a_r in terms of the Δq_i, with $1 \leq i \leq k$, metrics simultaneously, a new metric was defined, based on an aggregation operator, $Q_{rs}(\Delta q_1^{rs}, \Delta q_2^{rs}, \dots, \Delta q_k^{rs})$ or simply Q_{rs} for short, with $1 \leq r \leq l$; $1 \leq s \leq t$. In this case, the *arithmetic average* of the values of the metrics used was considered, as shown in equation (2). It is considered convenient to use an aggregate value of the values of the metrics used, to avoid biases that could occur when using a single metric.

$$Q_{rs}(\Delta q_1^{rs}, \Delta q_2^{rs}, \dots, \Delta q_k^{rs}) = \frac{1}{k}\sum_{i=1}^{k} \Delta q_i^{rs} \text{ ; with } \begin{matrix} 1 \leq s \leq t \\ 1 \leq r \leq l \end{matrix} \tag{2}$$

Thus, by ascendingly ordering the imputation methods by the values given by equation (2), we obtain the goodness-of-fit order of m_s, with $1 \leq s \leq t$, imputation methods used to impute the amputated Y data set in the combination a_r, with $1 \leq s \leq l$, with respect to the Δq_i, with $1 \leq i \leq k$, metrics simultaneously.

To evaluate the performance of m_s, with $1 \leq s \leq t$, imputation methods used to impute the amputated Y data sets in the a_r, with $1 \leq r \leq l$, combinations, i.e., *considering all amputation scenarios (all data sets considered)*, two criteria were used.

Criterion 1. It is considered a new metric $R_{si}(\Delta q_i^{rs}, \Delta q_i^{rs}, \dots, \Delta q_i^{rs})$ or simply R_{si} for short, with $1 \leq r \leq l$; $1 \leq i \leq k$ and $1 \leq s \leq t$, given by equation (3). This metric thus defined, allows to compute the *arithmetic average* of the values of the metric Δq_i $(Y^{a_r m_s})$, for the imputation method m_s used to impute all amputed data sets in the a_r combinations.

R_{si} is an average indicator of the performance of the s imputation method for all files amputated with different mechanisms and then imputed with the s method, considering one of the metrics Δq_i.

$$R_{si}(\Delta q_i \ , \Delta q_i \ , \dots, \Delta q_i \ ) = \frac{1}{l}\sum_{r=1}^{l} \Delta q_i(Y^{a_r m_s}); \text{ with} \begin{matrix} 1 \leq i \leq k \\ 1 \leq s \leq t \end{matrix} \tag{3}$$

Thus, by ascendingly ordering the imputation methods by the values given by equation (3), we obtain the order of goodness of fit of the m_s, with $1 \leq s \leq t$, imputation methods used to impute all amputated Y data sets in the a_r, with $1 \leq r \leq l$, combinations, with respect to the metric Δq_i, with $1 \leq i \leq k$.

Similarly, a new metric was defined, $T_s[Q_{rs}(\Delta q_1^{rs}, \Delta q_2^{rs}, \dots, \Delta q_k^{rs})]$ or simply T_s, as shown in equation (4), which allows to obtain the arithmetic average of the values of the metric Q_{rs} for the imputation method m_s, with $1 \leq s \leq t$, used to impute all amputed data sets in the a_r, with $1 \leq r \leq l$ combinations.

$$T_s[Q_{rs}(\Delta q_1^{rs}, \Delta q_2^{rs}, \dots, \Delta q_k^{rs})] \\ = \frac{1}{l}\sum_{r=1}^{l} Q_{rs}(\Delta q_1^{rs}, \Delta q_2^{rs}, \dots, \Delta q_k^{rs}) \text{ ; with } 1 \leq s \leq t \tag{4}$$

Ascendingly ordering the imputation methods by the values given by the first term of equation (4), we obtain the order of goodness of the m_s, with $1 \leq s \leq t$, imputation methods used to impute all amputated Y data sets in the a_r, with $1 \leq r \leq l$, combinations, with respect to the Δq_i metrics *simultaneously*, with $1 \leq i \leq k$.

T_s is an average indicator of the performance of the s imputation method for all files amputated with different mechanisms and then imputed with the s method, considering simultaneously all metrics Δq_i.

Criterion 2. It is considered the order of goodness of the imputation methods m_s, with $1 \leq s \leq t$, used to impute the amputed data set in the combination a_r, with $1 \leq r \leq l$, with respect to the metrics Δq_i^{rs}, with $1 \leq i \leq k$, and with respect to the metric Q_{rs}.

A score p_i^{rs} was assigned to the imputation method m_s, used to impute the amputated Y data set in the combination a_r, which comes *first in the order of goodness* of fit with respect to the values of the metric Δq_i^{rs} obtained using equation (1). Similarly, a score P_{rs} is assigned to the imputation method m_s, used to impute the amputated data set in the combination a_r, which comes first in the order of goodness of fit with respect to the values of the metric Q_{rs}.

The score was assigned according to the following criteria. If an imputation method m_s results first in the goodness-of-fit order, 1 (one) point is assigned to the method. If two imputation methods m_s and $m_{s'}$ tie for first place in the order of goodness of fit, ½ (half) point is assigned to each of them. If three imputation methods m_s, $m_{s'}$ and $m_{s''}$ tie for first place in the order of goodness of fit, each of them is assigned 1/3 (one third) of a point and, in general, if all t imputation methods tie for first place in the order of goodness of fit, each of them is assigned $1/t$ points.

Applying the above mentioned procedure, a new metric w_{si} was defined as shown in equation (5), as the score obtained by the imputation method m_s, considering the metric Δq_i. The value of w_{si} indicates the score obtained by the imputation method s for the metric Δq_i.

$$w_{si} = \sum_{r=1}^{l} p_i^{rs} \text{ ; with } \begin{matrix} 1 \leq i \leq k \\ 1 \leq s \leq t \end{matrix} \tag{5}$$

Sorting the imputation methods descendingly by the values given by equation (5), we obtain the order of goodness of fit of the m_s, with $1 \leq s \leq t$, imputation methods used to impute the amputated Y data sets in the a_r, with $1 \leq r \leq l$, combinations with respect to the w_{si} metric.

Similarly, a new metric W_s was defined as the score obtained by the imputation method m_s, considering the values of the metric P_{rs}, average score of all metrics Δq_i.

$$W_s = \sum_{r=1}^{l} P_{rs} \text{ ; with } 1 \leq s \leq t \tag{6}$$

Sorting the imputation methods descendingly by the values given by equation (6), we obtain the order of goodness of fit of the m_s, with $1 \leq s \leq t$, imputation methods used to impute the amputated Y data sets in the a_r, with $1 \leq r \leq l$, combinations with respect to the W_s metric.

Finally, a new metric G_s given by equation (7) was defined as the overall score obtained by each imputation method m_s considering all metrics, w_{si} and W_s.

$$G_s = \left(\sum_{i=1}^{k} w_{si}\right) + W_s; \text{ with } 1 \leq s \leq t \tag{7}$$

Sorting the imputation methods descendingly by the values given by equation (7), we obtain the order of goodness of fit of the m_s, with $1 \leq s \leq t$, imputation methods used to impute all amputated Y data sets in the a_r, with $1 \leq r \leq l$, combinations with respect to the G_s metric.

This metric is considered the global indicator of this proposal, although each of the summands of equation (7) separately could also be considered as proxy indicators.

IV. Results and Discussions

Table VII presents the values of the quality indicator metrics of the DMM obtained through the DMP using the original "Iris" dataset. These are *quality (Cal)*, *precision (accuracy) (Prec)* and *classification (Clas)*.

TABLE VII. Values of the Metrics for the Original Data Set (Own Elaboration)

	Quality (Cal)	Precision (Prec)	Classification (Clas)
Iris	0.884	0.972	0.796

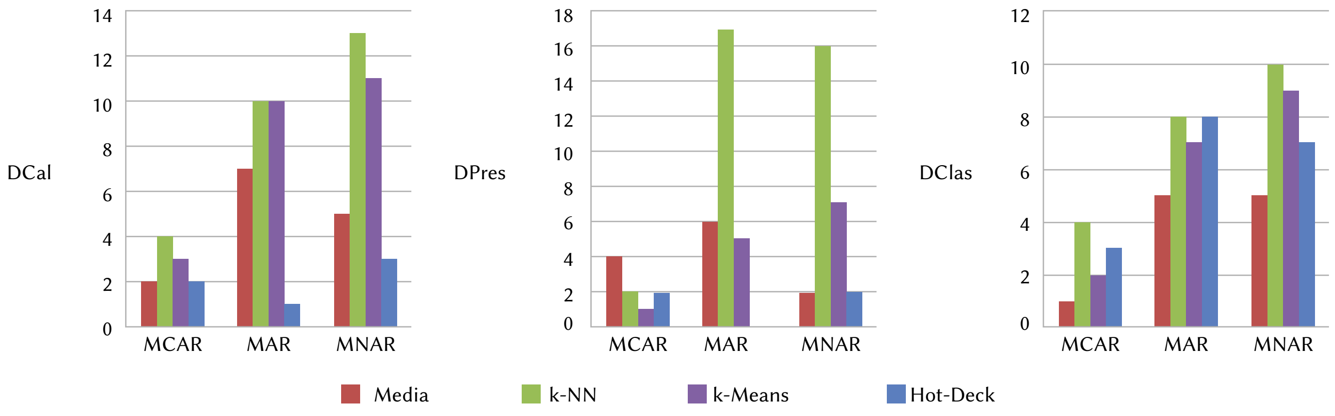


Fig. 1. First place according to MV mechanisms (Own elaboration).

Table VIII presents the values of the DMM quality indicator metrics obtained by DMP using the "Iris" datasets imputed by the Means, k-NN, k-Means and Hot-Deck imputation methods, after being amputated in each of the 63 combinations of mechanisms, patterns and MV percentages described in [11]. In total, for 63 amputated datasets, 252 imputed datasets were obtained (63 x 4). MR indicates the percentage of missing records.

Each row of Table VIII represents the characteristics of the amputated datasets and the value of each of the DMM goodness-of-fit indicator metrics obtained by DMP using the dataset imputed by Mean, k-NN, k-Means and Hot-Deck imputation methods after amputation.

Thus, for example, the *accuracy* value of the DMM obtained with the "Iris" data set imputed by the k-NN imputation method after having been amputated according to the MCAR assumption, in univariate pattern in 10% of the records is 0.967.

Table IX shows the values of the metrics *differences in absolute value* between the values of the DMM quality indicator metrics mentioned in Table VII and Table VIII, obtained using equation (1).

Thus, for example, the values of the *differences in absolute value* between the quality metrics (Δ*Cal*) for the "Iris" data sets imputed by Mean, k-NN, k-Means and Hot-Deck after amputation in the MCAR assumption, in univariate pattern on 10% of the records, are 0.007; 0.001; 0.004 and 0.008 respectively.

Sorting the preceding values in ascending order gives the k-NN, k-Means, Medians and Hot-Deck methods, ranked according to their order of goodness of fit for the relevant imputation method.

The results presented in Table IX for each of the metrics and the number of times each imputation method came first, second, third and fourth in the order of goodness of fit to impute each of the 63 amputated data sets are described below.

Regarding the *difference in absolute value* between the *quality* metrics (Δ*Cal*), the *Mean* imputation method came first, second, third and fourth 14, 1, 17 and 31 out of 63 times respectively. Also, of the 14 times it came first, it shared position with the k-NN method and in one with the k-NN and k-Means methods. In terms of the *absolute value difference* between the *precision* metrics (Δ*Prec*), the *Mean* imputation method came first, second, third and fourth 12, 5, 20 and 26 out of 63 times respectively. Finally, for the *absolute value differences* between the *classification* metrics (Δ*Clas*), the *Mean* imputation method came first, second, third and fourth 11, 19, 11 and 22 out of 63 times, respectively. Of the 12 times it came first in the order of goodness of fit, it was accompanied by the k-NN method once and the k-Means method once.

Regarding the *difference in absolute value* between the *quality* metrics (Δ*Cal*), the *k-NN* imputation method came first, second, third, and fourth in 27, 25, 8, and 3 of 63 times, respectively. Also, of the 27 times it came first in the goodness-of-fit order, in one it shared position with the Hot-Deck method and in three with the k-Means method. In terms of the *difference in absolute value* between the *precision* metrics (Δ*Prec*), the *k-NN* imputation method came first, second, third and fourth 35, 20, 5 and 3 times out of 63, respectively. Of the 35 times it came first, once it did so jointly with k-Means. Finally, for the *absolute value difference* between metric *classification* (Δ*Clas*), the *k-NN* imputation method came first, second, third and fourth 22, 13, 24 and 4 times out of 63, respectively. Of the 22 times it came first, four times it did so jointly with k-Means.

Regarding the *difference in absolute value* between the *quality* metrics (Δ*Cal*), the *k-Means* imputation method came first, second, third and fourth 24, 23, 14 and 2 out of 63 times, respectively. Likewise, of the 24 times it came first in the goodness-of-fit order, once it did so jointly with Mean and k-NN, 3 times with k-Means and once with Hot-Deck. In terms of the *difference in absolute value* between the *precision* metrics (Δ*Prec*), the *k-Means* imputation method came first, second, third and fourth 13, 31, 17 and 2 times out of 63, respectively. Of the 13 times it came first, once it did so jointly with k-NN. Finally, for the *absolute value difference* between the *classification* metrics Δ*Clas*), the *k-Means* imputation method came first, second, third and fourth 18, 14, 19 and 12 times out of 63, respectively. Of the 18 times it came first, once it did so jointly with Mean, twice with k-NN and once with Hot-Deck.

Regarding the *difference in absolute value* between the *quality* metrics (Δ*Cal*), the *Hot-Deck* imputation method came first, second, third and fourth in 6, 10, 20 and 27 out of 63 times, respectively. Also, of the 6 times it came first in the order of goodness of fit, it did so jointly with k-NN once and once with k-Means. In terms of the *absolute value difference* between the *precision* metrics (Δ*Prec*), the *Hot-Deck* imputation method came first, second, third and fourth 13, 31, 17 and 2 times out of 63, respectively. Finally, for the *absolute value difference* between *classification* metrics (Δ*Clas*), the *Hot-Deck* imputation method came first, second, third and fourth in 18, 18, 8 and 19 times out of 63 respectively. Of the 18 times it came first, once it did so jointly with k-NN and once with k-Means.

Fig. 1 presents the number of times that the Mean, k-NN, k-Means, and Hot-Deck imputation methods came first, with respect to each metric and under each of the three assumed *MV mechanisms.* It is clearly observed that the k-NN imputation method results first overall, except with respect to the Δ*Cal* and Δ*Clas* metrics under the MAR assumption where it ranks first with k-Means and with respect to the Δ*Prec* metric under the MCAR assumption where the first place is for Mean.

The number of times that the Mean, k-NN, k-Means and Hot-Deck imputation methods came first for each metric considering the three

TABLE VIII. Values of the Metrics for the Imputed Data Sets (Own Elaboration)

Amputation data set in the amputation combination				Imputation method used to impute the amputated dataset											
				Media			k-NN			k-Means			Hot-Deck		
Mechanism	**Type**	**Pattern**	**MR**	***Cal***	***Prec***	***Clas***	***Cal***	***Prec***	***Clas***	***Cal***	***Prec***	***Clas***	***Cal***	***Prec***	***Clas***
MCAR	**-**	**univa**	**0.1**	0.877	0.97	0.784	0.883	**0.967**	0.798	0.88	0.967	0.793	0.876	0.964	0.788
			0.15	0.877	0.971	0.783	0.889	0.972	0.806	0.888	0.979	0.796	0.868	0.949	0.786
			0.2	0.881	0.974	0.788	0.881	0.967	0.795	0.881	0.963	0.8	0.872	0.955	0.79
		multiva2	**0.1**	0.897	0.997	0.797	0.891	0.973	0.809	0.88	0.968	0.792	0.878	0.96	0.796
			0.15	0.818	0.872	0.763	0.896	0.985	0.806	0.898	0.99	0.806	0.88	0.965	0.796
			0.2	0.773	0.753	0.793	0.872	0.942	0.801	0.901	0.99	0.812	0.835	0.895	0.775
		multiva3	**0.1**	0.848	0.92	0.775	0.859	0.908	0.808	0.858	0.907	0.808	0.852	0.893	0.811
			0.15	0.753	0.718	0.788	0.86	0.916	0.803	0.855	0.902	0.808	0.879	0.978	0.781
			0.2	0.782	0.806	0.758	0.88	0.973	0.786	0.811	0.824	0.798	0.803	0.811	0.796
MAR	**LEFT**	**univa**	**0.1**	0.893	0.988	0.798	0.867	0.924	0.809	0.866	0.923	0.809	0.675	0.549	0.802
			0.15	0.892	0.993	0.79	0.874	0.943	0.805	0.799	0.792	0.806	0.79	0.782	0.798
			0.2	0.892	0.994	0.79	0.79	0.777	0.803	0.786	0.768	0.804	0.812	0.824	0.801
		multiva2	**0.1**	0.784	0.764	0.805	0.861	0.91	0.813	0.86	0.907	0.813	0.852	0.893	0.811
			0.15	0.811	0.828	0.793	0.862	0.912	0.813	0.861	0.909	0.813	0.618	0.464	0.772
			0.2	0.713	0.642	0.784	0.863	0.919	0.808	0.865	0.919	0.812	0.863	0.908	0.818
		multiva3	**0.1**	0.742	0.716	0.768	0.88	0.991	0.768	0.859	0.912	0.806	0.817	0.826	0.808
			0.15	0.787	0.798	0.777	0.876	0.988	0.764	0.89	0.973	0.806	0.842	0.876	0.808
			0.2	0.815	0.877	0.753	0.812	0.86	0.764	0.893	0.973	0.813	0.842	0.93	0.755
	MID	**univa**	**0.1**	0.826	0.866	0.786	0.896	0.989	0.803	0.861	0.912	0.81	0.794	0.814	0.775
			0.15	0.881	0.972	0.79	0.89	0.981	0.799	0.863	0.917	0.81	0.593	0.435	0.752
			0.2	0.879	0.955	0.804	0.89	0.981	0.799	0.895	0.986	0.805	0.835	0.906	0.764
		multiva2	**0.1**	0.795	0.784	0.806	0.905	1	0.81	0.86	0.909	0.812	0.853	0.908	0.798
			0.15	0.757	0.707	0.808	0.882	0.96	0.803	0.9	0.991	0.808	0.594	0.453	0.734
			0.2	0.753	0.697	0.808	0.884	0.965	0.803	0.883	0.956	0.81	0.799	0.797	0.802
		multiva3	**0.1**	0.812	0.873	0.752	0.873	0.979	0.768	0.893	0.976	0.81	0.722	0.649	0.796
			0.15	0.839	0.943	0.736	0.872	0.976	0.768	0.894	0.988	0.8	0.841	0.869	0.813
			0.2	0.802	0.873	0.731	0.856	0.965	0.746	0.894	0.983	0.805	0.731	0.709	0.753
	RIGHT	**univa**	**0.1**	0.793	0.794	0.791	0.885	0.983	0.787	0.853	0.889	0.818	0.863	0.936	0.79
			0.15	0.878	0.956	0.8	0.89	0.981	0.799	0.861	0.922	0.8	0.576	0.411	0.74
			0.2	0.884	0.965	0.803	0.89	0.981	0.799	0.859	0.904	0.815	0.744	0.727	0.76
		multiva2	**0.1**	0.787	0.765	0.81	0.9	0.992	0.808	0.896	0.982	0.81	0.893	0.998	0.788
			0.15	0.773	0.742	0.804	0.887	0.97	0.803	0.895	0.982	0.808	0.822	0.854	0.791
			0.2	0.744	0.695	0.793	0.887	0.978	0.797	0.882	0.955	0.81	0.644	0.553	0.734
		multiva3	**0.1**	0.855	0.976	0.734	0.818	0.891	0.745	0.898	0.998	0.798	0.858	0.912	0.803
			0.15	0.816	0.898	0.734	0.84	0.936	0.745	0.861	0.917	0.805	0.819	0.839	0.8
			0.2	0.785	0.865	0.705	0.858	0.976	0.741	0.894	0.999	0.789	0.858	0.928	0.787
MNAR	**LEFT**	**univa**	**0.1**	0.893	0.988	0.798	0.867	0.924	0.809	0.867	0.923	0.811	0.675	0.549	0.802
			0.15	0.889	0.988	0.79	0.89	0.981	0.799	0.87	0.934	0.806	0.704	0.62	0.789
			0.2	0.889	0.988	0.79	0.89	0.981	0.799	0.873	0.94	0.806	0.846	0.886	0.805
		multiva2	**0.1**	0.789	0.771	0.806	0.861	0.91	0.813	0.861	0.908	0.813	0.856	0.901	0.811
			0.15	0.823	0.852	0.793	0.862	0.91	0.813	0.86	0.902	0.818	0.767	0.757	0.776
			0.2	0.726	0.673	0.779	0.861	0.91	0.811	0.861	0.909	0.813	0.859	0.907	0.811
		multiva3	**0.1**	0.714	0.662	0.767	0.871	0.956	0.786	0.859	0.911	0.806	0.852	0.897	0.808
			0.15	0.834	0.9	0.769	0.872	0.977	0.766	0.857	0.902	0.812	0.849	0.886	0.812
			0.2	0.766	0.786	0.747	0.825	0.892	0.757	0.896	0.908	0.812	0.855	0.929	0.781
	MID	**univa**	**0.1**	0.753	0.721	0.784	0.89	0.981	0.799	0.863	0.92	0.806	0.878	0.985	0.771
			0.15	0.883	0.975	0.79	0.89	0.981	0.799	0.863	0.92	0.806	0.55	0.351	0.75
			0.2	0.888	0.987	0.788	0.89	0.981	0.799	0.872	0.938	0.806	0.796	0.845	0.747
		multiva2	**0.1**	0.814	0.825	0.803	0.856	0.899	0.813	0.852	0.887	0.816	0.732	0.667	0.797
			0.15	0.757	0.707	0.808	0.882	0.96	0.803	0.9	0.991	0.808	0.594	0.453	0.734
			0.2	0.753	0.697	0.808	0.884	0.965	0.803	0.883	0.956	0.81	0.799	0.797	0.802
		multiva3	**0.1**	0.852	0.952	0.752	0.875	0.982	0.768	0.894	0.978	0.81	0.585	0.402	0.768
			0.15	0.829	0.917	0.741	0.848	0.95	0.746	0.884	0.971	0.796	0.847	0.906	0.788
			0.2	0.802	0.872	0.731	0.846	0.946	0.746	0.895	0.982	0.808	0.748	0.734	0.763
	RIGHT	**univa**	**0.1**	0.774	0.755	0.792	0.874	0.967	0.782	0.894	0.988	0.8	0.69	0.607	0.774
			0.15	0.688	0.612	0.764	0.739	0.719	0.759	0.892	0.985	0.8	0.838	0.907	0.769
			0.2	0.893	0.991	0.795	0.89	0.981	0.799	0.854	0.887	0.82	0.865	0.969	0.76
		multiva2	**0.1**	0.787	0.765	0.81	0.9	0.992	0.808	0.899	0.988	0.81	0.893	0.998	0.788
			0.15	0.773	0.742	0.804	0.887	0.97	0.803	0.883	0.952	0.815	0.801	0.828	0.775
			0.2	0.744	0.695	0.763	0.887	0.978	0.797	0.88	0.95	0.81	0.644	0.533	0.734
		multiva3	**0.1**	0.511	0.294	0.728	0.567	0.368	0.766	0.897	0.996	0.798	0.871	0.942	0.8
			0.15	0.658	0.591	0.724	0.82	0.893	0.746	0.889	0.984	0.795	0.894	0.99	0.798
			0.2	0.671	0.63	0.713	0.856	0.96	0.751	0.857	0.91	0.803	0.809	0.817	0.801

TABLE IX. Value of the Metrics Differences in Absolute Value (Own Elaboration).

Amputation data set in the amputation combination				**Imputation method**											
				Medias			**k-NN**			**k-Means**			**Hot-Deck**		
Mechanism	**Type**	**Pattern**	**MR**	Δ*Cal*	Δ*Prec*	Δ*Clas*	Δ*Cal*	Δ*Prec*	Δ*Clas*	Δ*Cal*	Δ*Prec*	Δ*Clas*	Δ*Cal*	Δ*Prec*	Δ*Clas*
MCAR	**-**	**univa**	**0.1**	**0.007**	0.002	0.012	**0.001**	0.005	0.002	**0.004**	0.005	0.003	**0.008**	0.008	0.008
			0.15	0.007	0.001	0.013	0.005	0.000	0.010	0.004	0.007	0.000	0.016	0.023	0.010
			0.2	0.003	0.002	0.008	0.003	0.005	0.001	0.003	0.009	0.004	0.012	0.017	0.006
		multiva2	**0.1**	0.013	0.025	0.001	0.007	0.001	0.013	0.004	0.004	0.004	0.006	0.012	0.000
			0.15	0.066	0.100	0.033	0.012	0.013	0.010	0.014	0.018	0.010	0.004	0.007	0.000
			0.2	0.111	0.219	0.003	0.012	0.030	0.005	0.017	0.018	0.016	0.049	0.077	0.021
		multiva3	**0.1**	0.036	0.052	0.021	0.025	0.064	0.012	0.026	0.065	0.012	0.032	0.079	0.015
			0.15	0.131	0.254	0.008	0.024	0.056	0.007	0.029	0.070	0.012	0.005	0.006	0.015
			0.2	0.014	0.010	0.038	0.084	0.177	0.010	0.073	0.148	0.002	0.081	0.161	0.000
MAR	**LEFT**	**univa**	**0.1**	0.009	0.016	0.002	0.017	0.048	0.013	0.018	0.049	0.013	0.209	0.423	0.006
			0.15	0.008	0.021	0.006	0.010	0.029	0.009	0.085	0.180	0.010	0.094	0.190	0.002
			0.2	0.008	0.022	0.006	0.094	0.195	0.007	0.098	0.204	0.008	0.072	0.148	0.005
		multiva2	**0.1**	0.100	0.208	0.009	0.023	0.062	0.017	0.024	0.065	0.017	0.032	0.079	0.015
			0.15	0.073	0.144	0.003	0.022	0.060	0.017	0.023	0.063	0.017	0.266	0.508	0.024
			0.2	0.171	0.330	0.012	0.021	0.053	0.012	0.019	0.053	0.016	0.021	0.064	0.022
		multiva3	**0.1**	0.142	0.256	0.028	0.004	0.019	0.028	0.025	0.060	0.010	0.067	0.146	0.012
			0.15	0.097	0.174	0.019	0.008	0.016	0.032	0.006	0.001	0.010	0.042	0.096	0.012
			0.2	0.069	0.095	0.043	0.072	0.112	0.032	0.009	0.001	0.017	0.042	0.042	0.041
	MID	**univa**	**0.1**	0.058	0.106	0.010	0.012	0.017	0.007	0.023	0.060	0.014	0.090	0.158	0.021
			0.15	0.003	0.000	0.006	0.006	0.009	0.003	0.021	0.055	0.014	0.291	0.537	0.044
			0.2	0.005	0.017	0.008	0.006	0.009	0.003	0.011	0.014	0.009	0.049	0.066	0.032
		multiva2	**0.1**	0.089	0.188	0.010	0.021	0.028	0.014	0.024	0.063	0.016	0.031	0.064	0.002
			0.15	0.127	0.265	0.012	0.002	0.012	0.007	0.016	0.019	0.012	0.290	0.519	0.062
			0.2	0.131	0.275	0.012	0.000	0.007	0.007	0.001	0.016	0.014	0.085	0.175	0.006
		multiva3	**0.1**	0.072	0.099	0.044	0.011	0.007	0.028	0.009	0.004	0.014	0.162	0.323	0.000
			0.15	0.045	0.029	0.060	0.012	0.004	0.028	0.010	0.016	0.004	0.043	0.103	0.017
			0.2	0.082	0.099	0.065	0.028	0.007	0.050	0.010	0.011	0.009	0.153	0.263	0.043
	RIGHT	**univa**	**0.1**	0.091	0.178	0.005	0.001	0.011	0.009	0.031	0.083	0.022	0.021	0.036	0.006
			0.15	0.006	0.016	0.004	0.006	0.009	0.003	0.023	0.050	0.004	0.308	0.561	0.056
			0.2	0.000	0.007	0.007	0.006	0.009	0.003	0.025	0.068	0.019	0.140	0.245	0.036
		multiva2	**0.1**	0.097	0.207	0.014	0.016	0.020	0.012	0.012	0.010	0.014	0.009	0.026	0.008
			0.15	0.111	0.230	0.008	0.003	0.002	0.007	0.011	0.010	0.012	0.062	0.118	0.005
			0.2	0.140	0.277	0.003	0.003	0.006	0.001	0.002	0.017	0.014	0.240	0.419	0.062
		multiva3	**0.1**	0.029	0.004	0.062	0.066	0.081	0.051	0.014	0.026	0.002	0.026	0.060	0.007
			0.15	0.068	0.074	0.062	0.044	0.036	0.051	0.023	0.055	0.009	0.065	0.133	0.004
			0.2	0.099	0.107	0.091	0.026	0.004	0.055	0.010	0.027	0.007	0.026	0.044	0.009
MNAR	**LEFT**	**univa**	**0.1**	0.009	0.016	0.002	0.017	0.048	0.013	0.017	0.049	0.015	0.209	0.423	0.006
			0.15	0.005	0.016	0.006	0.006	0.009	0.003	0.014	0.038	0.010	0.180	0.352	0.007
			0.2	0.005	0.016	0.006	0.006	0.009	0.003	0.011	0.032	0.010	0.038	0.086	0.009
		multiva2	**0.1**	0.095	0.201	0.010	0.023	0.062	0.017	0.023	0.064	0.017	0.028	0.071	0.015
			0.15	0.061	0.120	0.003	0.022	0.062	0.017	0.024	0.070	0.022	0.117	0.215	0.020
			0.2	0.158	0.299	0.017	0.023	0.062	0.015	0.023	0.063	0.017	0.025	0.065	0.015
		multiva3	**0.1**	0.170	0.310	0.029	0.013	0.016	0.010	0.025	0.061	0.010	0.032	0.075	0.012
			0.15	0.050	0.072	0.027	0.012	0.005	0.030	0.027	0.070	0.016	0.035	0.086	0.016
			0.2	0.118	0.186	0.049	0.059	0.080	0.039	0.012	0.064	0.016	0.029	0.043	0.015
	MID	**univa**	**0.1**	0.131	0.251	0.012	0.006	0.009	0.003	0.021	0.052	0.010	0.006	0.013	0.025
			0.15	0.001	0.003	0.006	0.006	0.009	0.003	0.021	0.052	0.010	0.334	0.621	0.046
			0.2	0.004	0.015	0.008	0.006	0.009	0.003	0.012	0.034	0.010	0.088	0.127	0.049
		multiva2	**0.1**	0.070	0.147	0.007	0.028	0.073	0.017	0.032	0.085	0.020	0.152	0.305	0.001
			0.15	0.127	0.265	0.012	0.002	0.012	0.007	0.016	0.019	0.012	0.290	0.519	0.062
			0.2	0.131	0.275	0.012	0.000	0.007	0.007	0.001	0.016	0.014	0.085	0.175	0.006
		multiva3	**0.1**	0.032	0.020	0.044	0.009	0.010	0.028	0.010	0.006	0.014	0.299	0.570	0.028
			0.15	0.055	0.055	0.055	0.036	0.022	0.050	0.000	0.001	0.000	0.037	0.066	0.008
			0.2	0.082	0.100	0.065	0.038	0.026	0.050	0.011	0.010	0.012	0.136	0.238	0.033
	RIGHT	**univa**	**0.1**	0.110	0.217	0.004	0.010	0.005	0.014	0.010	0.016	0.004	0.194	0.365	0.022
			0.15	0.196	0.360	0.032	0.145	0.253	0.037	0.008	0.013	0.004	0.046	0.065	0.027
			0.2	0.009	0.019	0.001	0.006	0.009	0.003	0.030	0.085	0.024	0.019	0.003	0.036
		multiva2	**0.1**	0.097	0.207	0.014	0.016	0.020	0.012	0.015	0.016	0.014	0.009	0.026	0.008
			0.15	0.111	0.230	0.008	0.003	0.002	0.007	0.001	0.020	0.019	0.083	0.144	0.021
			0.2	0.140	0.277	0.033	0.003	0.006	0.001	0.004	0.022	0.014	0.240	0.439	0.062
		multiva3	**0.1**	0.373	0.678	0.068	0.317	0.604	0.030	0.013	0.024	0.002	0.013	0.030	0.004
			0.15	0.226	0.381	0.072	0.064	0.079	0.050	0.005	0.012	0.001	0.010	0.018	0.002
			0.2	0.213	0.342	0.083	0.028	0.012	0.045	0.027	0.062	0.007	0.075	0.155	0.005

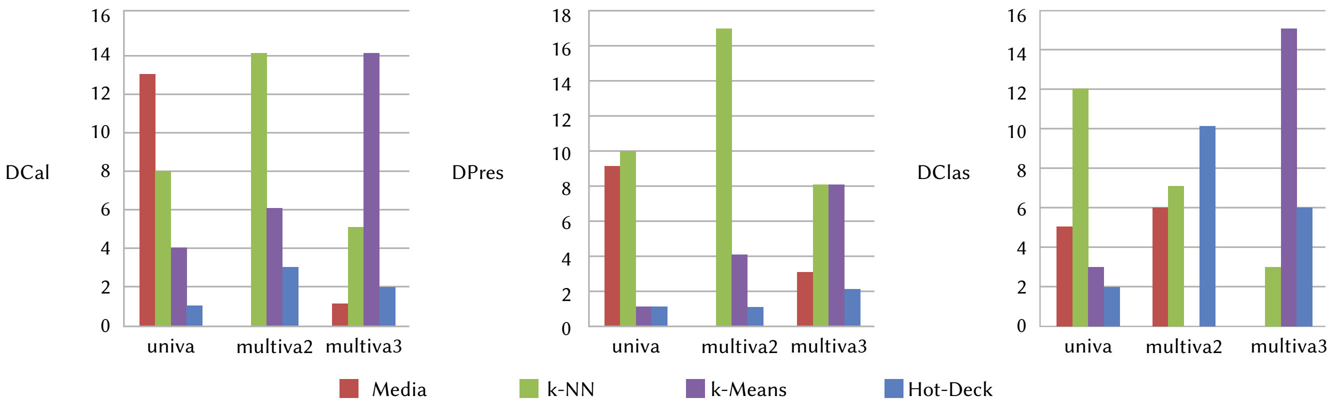


Fig. 2. First place according to MV patterns (Own elaboration).

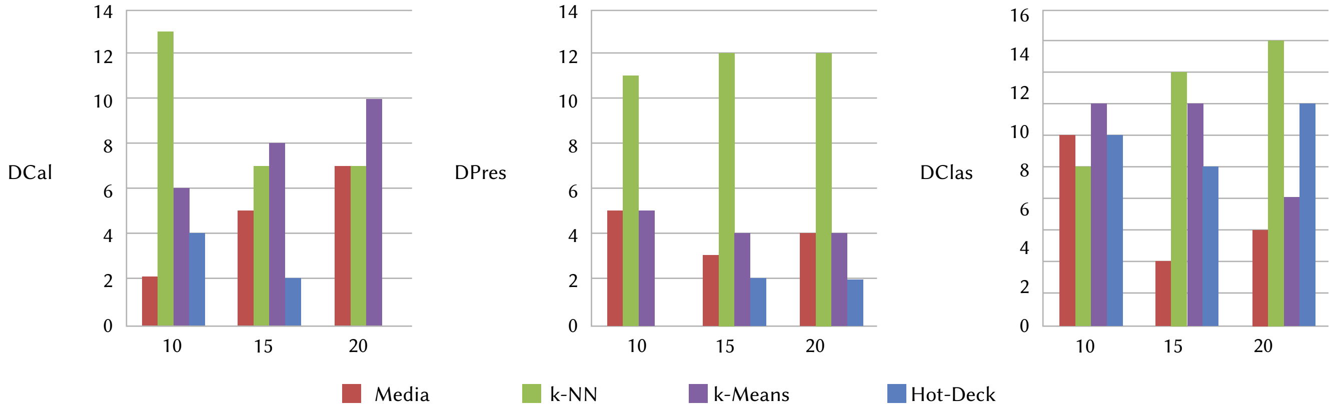


Fig. 3. First place according to percentage of MV (Own elaboration).

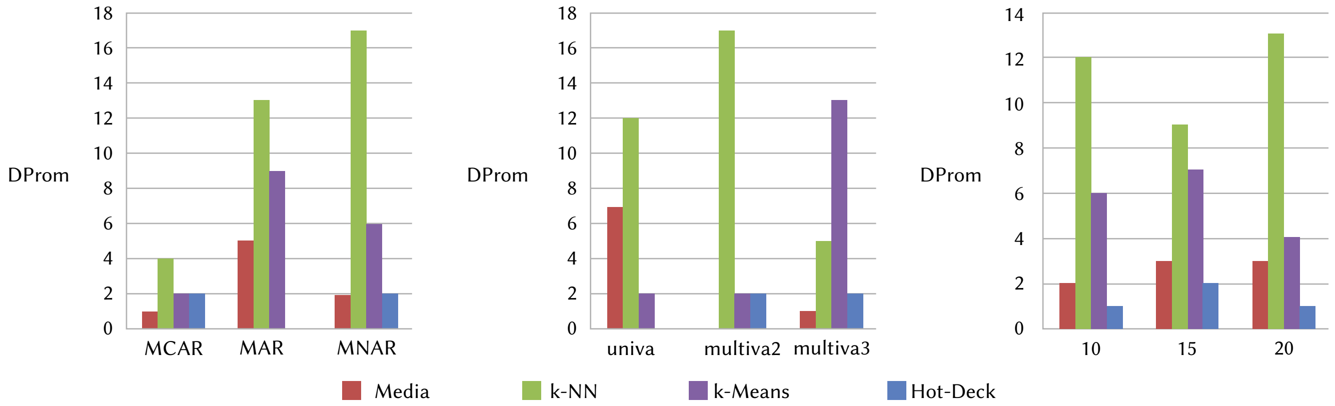


Fig. 4. First place with respect to the arithmetic average metric (Own elaboration).

MV patterns is presented in Fig. 2. The graphs show a clear dispute for first place between the k-NN and k-Means methods. Regarding the Δ*Cal* metric, the Mean imputation method clearly results in the first place when dealing with a univariate pattern. However, in the case of a simple multivariate pattern k-NN comes first; something similar happens with the complex multivariate pattern where k-Means comes first. Concerning Δ*Prec*, the first place is for k-NN for both the univariate and simple multivariate pattern, however, it shares the first place with k-Means in the case of a complex multivariate pattern. Finally, regarding *Clas*, the results are mixed, k-NN came first in the case of a univariate pattern, Hot-Deck in the case of a simple multivariate one and k-Means in the case of complex multivariate pattern.

Finally, the number of times that the Mean, k-NN, k-Means and Hot-Deck imputation methods came first, with respect to each metric and considering the different *MV percentages* are shown in Fig. 3. k-NN comes first with respect to Δ*Cal* for an MV percentage of 10% while for 15% and 20% k-Means comes first. With respect to Δ*Prec*, it is clearly observed that in all cases k-NN comes out first. Finally, with respect to Δ*Clas*, k-Means came first for 10% while k-NN came first for 15% and 20%.

In Table X, the values of the metrics obtained using equation (2), i.e., the *arithmetic average* of the metric values Δ*Cal*, Δ*Prec* and Δ*Clas*, indicated in Table IX, for each imputation method m_s used to impute the amputed data set in the combination a_r, are presented.

By sorting the imputation methods in ascending order by the values of this metric, we obtain the order of goodness of fit of the Medias, k-NN, k-Means and Hot-Deck imputation methods used to impute the "Iris" data set in each of the 63 amputation combinations.

TABLE X. Arithmetic Average Metric Values of ΔCal, $\Delta Prec$ and $\Delta Clas$ (Own Elaboration)

Amputation combination				**Imputation method**			
				Medias	**k-NN**	**k-Means**	**Hot-Deck**
Mechanism	**Type**	**Pattern**	**MR**	**Average (ΔQ)**	**Average (ΔQ)**	**Average (ΔQ)**	**Average (ΔQ)**
MCAR	**-**	**univa**	**0.1**	0.007	0.003	0.004	0.008
MCAR	**-**		**0.15**	0.007	0.005	0.004	0.016
MCAR	**-**		**0.2**	0.004	0.003	0.005	0.012
MCAR	**-**	**multiva2**	**0.1**	0.013	0.007	0.004	0.006
MCAR	**-**		**0.15**	0.066	0.012	0.014	0.004
MCAR	**-**		**0.2**	0.111	0.016	0.017	0.049
MCAR	**-**	**multiva3**	**0.1**	0.036	0.034	0.034	0.042
MCAR	**-**		**0.15**	0.131	0.029	0.037	0.009
MCAR	**-**		**0.2**	0.021	0.090	0.074	0.081
MAR	**LEFT**	**univa**	**0.1**	0.009	0.026	0.027	0.213
MAR	**LEFT**		**0.15**	0.012	0.016	0.092	0.095
MAR	**LEFT**		**0.2**	0.012	0.099	0.103	0.075
MAR	**LEFT**	**multiva2**	**0.1**	0.106	0.034	0.035	0.042
MAR	**LEFT**		**0.15**	0.073	0.033	0.034	0.266
MAR	**LEFT**		**0.2**	0.171	0.029	0.029	0.036
MAR	**LEFT**	**multiva3**	**0.1**	0.142	0.017	0.032	0.075
MAR	**LEFT**		**0.15**	0.097	0.019	0.006	0.050
MAR	**LEFT**		**0.2**	0.069	0.072	0.009	0.042
MAR	**MID**	**univa**	**0.1**	0.058	0.012	0.032	0.090
MAR	**MID**		**0.15**	0.003	0.006	0.030	0.291
MAR	**MID**		**0.2**	0.010	0.006	0.011	0.049
MAR	**MID**	**multiva2**	**0.1**	0.096	0.021	0.034	0.032
MAR	**MID**		**0.15**	0.135	0.007	0.016	0.290
MAR	**MID**		**0.2**	0.139	0.005	0.010	0.089
MAR	**MID**	**multiva3**	**0.1**	0.072	0.015	0.009	0.162
MAR	**MID**		**0.15**	0.045	0.015	0.010	0.054
MAR	**MID**		**0.2**	0.082	0.028	0.010	0.153
MAR	**RIGHT**	**univa**	**0.1**	0.091	0.007	0.045	0.021
MAR	**RIGHT**		**0.15**	0.009	0.006	0.026	0.308
MAR	**RIGHT**		**0.2**	0.005	0.006	0.037	0.140
MAR	**RIGHT**	**multiva2**	**0.1**	0.106	0.016	0.012	0.014
MAR	**RIGHT**		**0.15**	0.116	0.004	0.011	0.062
MAR	**RIGHT**		**0.2**	0.140	0.003	0.011	0.240
MAR	**RIGHT**	**multiva3**	**0.1**	0.032	0.066	0.014	0.031
MAR	**RIGHT**		**0.15**	0.068	0.044	0.029	0.067
MAR	**RIGHT**		**0.2**	0.099	0.028	0.015	0.026
MNAR	**LEFT**	**univa**	**0.1**	0.009	0.026	0.027	0.213
MNAR	**LEFT**		**0.15**	0.009	0.006	0.021	0.180
MNAR	**LEFT**		**0.2**	0.009	0.006	0.018	0.044
MNAR	**LEFT**	**multiva2**	**0.1**	0.102	0.034	0.035	0.038
MNAR	**LEFT**		**0.15**	0.061	0.034	0.039	0.117
MNAR	**LEFT**		**0.2**	0.158	0.033	0.034	0.035
MNAR	**LEFT**	**multiva3**	**0.1**	0.170	0.013	0.032	0.040
MNAR	**LEFT**		**0.15**	0.050	0.016	0.038	0.046
MNAR	**LEFT**		**0.2**	0.118	0.059	0.031	0.029
MNAR	**MID**	**univa**	**0.1**	0.131	0.006	0.028	0.015
MNAR	**MID**		**0.15**	0.003	0.006	0.028	0.334
MNAR	**MID**		**0.2**	0.009	0.006	0.019	0.088
MNAR	**MID**	**multiva2**	**0.1**	0.075	0.039	0.046	0.153
MNAR	**MID**		**0.15**	0.135	0.007	0.016	0.290
MNAR	**MID**		**0.2**	0.139	0.005	0.010	0.089
MNAR	**MID**	**multiva3**	**0.1**	0.032	0.016	0.010	0.299
MNAR	**MID**		**0.15**	0.055	0.036	0.000	0.037
MNAR	**MID**		**0.2**	0.082	0.038	0.011	0.136
MNAR	**RIGHT**	**univa**	**0.1**	0.110	0.010	0.010	0.194
MNAR	**RIGHT**		**0.15**	0.196	0.145	0.008	0.046
MNAR	**RIGHT**		**0.2**	0.010	0.006	0.046	0.019
MNAR	**RIGHT**	**multiva2**	**0.1**	0.106	0.016	0.015	0.014
MNAR	**RIGHT**		**0.15**	0.116	0.004	0.013	0.083
MNAR	**RIGHT**		**0.2**	0.150	0.003	0.013	0.247
MNAR	**RIGHT**	**multiva3**	**0.1**	0.373	0.317	0.013	0.016
MNAR	**RIGHT**		**0.15**	0.226	0.064	0.006	0.010
MNAR	**RIGHT**		**0.2**	0.213	0.028	0.032	0.078

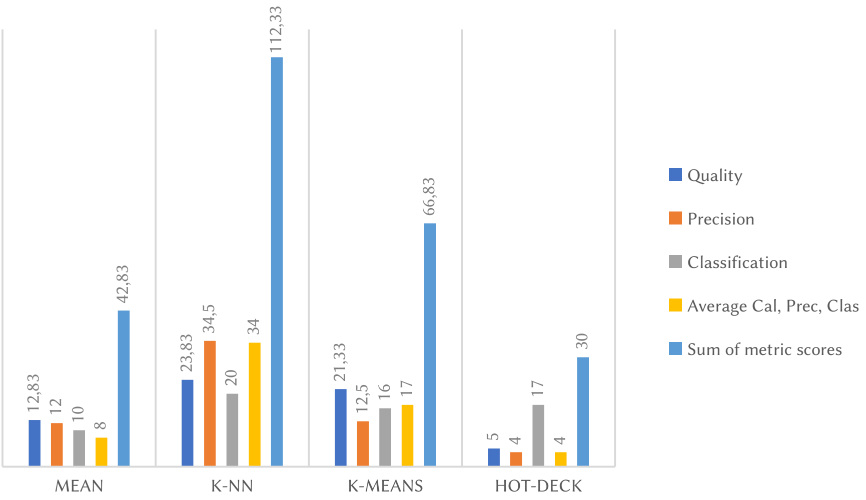


Fig. 5. Overall scores obtained by the imputation methods according to the metrics used (Own elaboration).

For example, by sorting the imputation methods in ascending order by the values indicated in the first row, we obtain the order of goodness of the imputation methods Medias, k-NN, k-Means and Hot-Deck used to impute the "Iris" data set after the original "Iris" data set was amputated according to the MCAR mechanism/assumption, in a univariate pattern on 10% of the records.

The results presented in Table X for this metric and the number of times each imputation method came first, second, third and fourth in the order of goodness of fit to impute each of the 63 amputated data sets are summarized below.

The *Mean* imputation method came first, second, third and fourth in 8, 7, 19 and 29 times out of 63, respectively. Likewise, *k-NN* ranked first, second, third and fourth 34, 17, 9 and 3 out of 63 times. The *k-Means* method came first, second, third and fourth 18, 14, 19 and 12 times out of 63, and finally, *Hot-Deck* came first, second, third and fourth 4, 12, 18 and 29 times out of 63, respectively.

Fig. 4 shows the number of times that the Mean, k-NN, k-Means and Hot-Deck methods came *first* in order of goodness of fit with respect to the *arithmetic average* aggregation operator metric and considering MV mechanisms, patterns and percentages. Clearly, the k-NN method came out first in all cases, except in the case of a complex multivariate MV pattern, where the k-Means method came out first.

Table XI shows the results obtained by applying equations (3) and (4), defined in *Criterion 1*, to the values obtained in Tables IX and X, i.e., the *arithmetic average* of the values of the quality, precision, classification, and aggregate metrics obtained by each imputation method.

By ascending the values in Table XI, the imputation methods were obtained for each metric, according to their order of goodness of fit.

TABLE XI. Values of the Arithmetic Average Metrics (Own Elaboration)

Imputation Method	Metrics			
	Pro. Δ*Cal*	**Pro. Δ*Pre***	**Pro. Δ*Clas***	**Pro. Met. Agr.**
Media	0.081	0.146	0.023	0.083
k-NN	0.026	0.044	0.017	0.029
k-Means	0.019	0.043	0.011	0.024
Hot-Deck	0.095	0.178	0.019	0.097

Regarding the *arithmetic average* of the values of the Δ*Cal* metric (Pro. Δ*Cal*), the k-Means, k-NN, Mean and Hot-Deck methods resulted according to their order of goodness of fit. Similarly, considering the *arithmetic average* of the values of the Δ*Prec* metric (Pro. Δ*Prec*), the k-Means, k-NN, Mean and Hot-Deck methods were obtained, according to their order of goodness. However, considering *arithmetic average* of the values of the Δ*Clas* metric (Pro. Δ*Clas*), the k-Means, k-NN, Hot-Deck and Mean methods resulted. Finally, with respect to the *arithmetic average* of the aggregate metric values (Pro. *Met. Agr.*), the k-Means, k-NN, Mean and Hot-Deck methods resulted according to their order of goodness of fit.

Table XII presents the scores obtained by the imputation methods that came first in the order of goodness of fit with respect to the metrics Δ*Cal*, Δ*Prec* and Δ*Clas* considering the values obtained using equation (1) and presented in Table IX, i.e., considering *Criterion 2.*

Thus, for example, considering the order of goodness of IM given by the value of the Δ*Cal*, Δ*Prec* and Δ*Clas* metrics in Table IX, with respect to the Δ*Cal* metric, one point was assigned to the k-NN IM used to impute the amputated "Iris" dataset according to the MCAR mechanism, in univariate pattern, in 10% of the records; similarly, with respect to the Δ*Prec* metric, the Mean imputation method scored one point when imputing the amputated "Iris" dataset according to the MCAR mechanism, in univariate pattern, in 10% of the records.

Similarly, with respect to the Δ*Cal* metric, 0.33 points were assigned to the IM by Mean, k-NN and k-Means used to impute the amputated "Iris" dataset according to the MCAR mechanism, in univariate pattern, in 20% of the records.

Similarly, with respect to the Δ*Clas* metric, 0.5 points were assigned to the MI k-Means and Hot-Deck used to impute the amputated "Iris" dataset according to the MCAR mechanism, in complex multivariate pattern, in 10% of the records.

Table XIII presents the scores obtained, considering *Criterion 2*, by the imputation methods that resulted first in the order of goodness of fit with respect to the aggregate metric considering the values obtained by equation (2) (metric Δ*Q* average of the metrics Δ*Cal*, Δ*Prec* and Δ*Clas*) and systematized in Table X.

Thus, for example, considering the order of goodness of IM given by the value of the ΔQ metric in Table X, a point was assigned to the k-NN IM used to impute the amputated "Iris" data set according to the MCAR mechanism, in univariate pattern, in 10% of the records.

Finally, Table XIV summarizes the score obtained by each IM for each metric, resulting from applying equations (5) and (6) to the data in Tables XII and XIII, and the overall score obtained by each imputation method, resulting from applying equation (7) to Table XIV.

TABLE XII. Scores Obtained for Each Metric (Own Elaboration)

Characteristics of Amputated Datasets				Scores obtained by each Imputation Method for each metric											
				Media			k-NN			k-Means			Hot-Deck		
Mechanism	Type	Pattern	MR	$p_1(\Delta Cal)$	$p_2(\Delta Prec)$	$p_3(\Delta Clas)$	$p_1(\Delta Cal)$	$p_2(\Delta Prec)$	$p_3(\Delta Clas)$	$p_1(\Delta Cal)$	$p_2(\Delta Prec)$	$p_3(\Delta Clas)$	$p_1(\Delta Cal)$	$p_2(\Delta Prec)$	$p_3(\Delta Clas)$
MCAR			0.1		1.00		**1.00**		1.00						
MCAR		univa	0.15					1.00		1.00		1.00			
MCAR			0.2	**0.33**	1.00		**0.33**		1.00	**0.33**					
MCAR			0.1					1.00		1.00					1.00
MCAR		multiva2	0.15										1.00	1.00	1.00
MCAR			0.2			1.00	1.00				1.00				
MCAR			0.1		1.00		1.00		**0.50**			**0.50**			
MCAR		multiva3	0.15						1.00				1.00	1.00	
MCAR			0.2	1.00	1.00										1.00
MAR	LEFT		0.1	1.00	1.00	1.00									
MAR	LEFT	univa	0.15	1.00	1.00										1.00
MAR	LEFT		0.2	1.00	1.00										1.00
MAR	LEFT		0.1			1.00	1.00	1.00							
MAR	LEFT	multiva2	0.15			1.00	1.00	1.00							
MAR	LEFT		0.2			0.50		0.50	0.50	1.00	0.50				
MAR	LEFT		0.1				1.00	1.00				1.00			
MAR	LEFT	multiva3	0.15							1.00	1.00	1.00			
MAR	LEFT		0.2							1.00	1.00	1.00			
MAR	MID		0.1				1.00	1.00	1.00						
MAR	MID	univa	0.15	1.00	1.00				1.00						
MAR	MID		0.2	1.00				1.00	1.00						
MAR	MID		0.1				1.00	1.00							1.00
MAR	MID	multiva2	0.15				1.00	1.00	1.00						
MAR	MID		0.2				1.00	1.00							1.00
MAR	MID		0.1							1.00	1.00				1.00
MAR	MID	multiva3	0.15					1.00		1.00		1.00			
MAR	MID		0.2					1.00		1.00		1.00			
MAR	RIGHT		0.1			1.00	1.00	1.00							
MAR	RIGHT	univa	0.15	0.50			0.50	1.00	1.00						
MAR	RIGHT		0.2	1.00	1.00				1.00						
MAR	RIGHT		0.1								1.00		1.00		1.00
MAR	RIGHT	multiva2	0.15				1.00	1.00							1.00
MAR	RIGHT		0.2					1.00	1.00	1.00					
MAR	RIGHT		0.1		1.00					1.00		1.00			
MAR	RIGHT	multiva3	0.15					1.00		1.00					1.00
MAR	RIGHT		0.2					1.00		1.00		1.00			
MNAR	LEFT		0.1	1.00	1.00	1.00									
MNAR	LEFT	univa	0.15	1.00				1.00	1.00						
MNAR	LEFT		0.2	1.00				1.00	1.00						
MNAR	LEFT		0.1			1.00	0.50	1.00		0.50					
MNAR	LEFT	multiva2	0.15			1.00	1.00	1.00							
MNAR	LEFT		0.2				0.50	1.00	0.50	0.50					0.50
MNAR	LEFT		0.1				1.00	1.00	0.50			0.50			
MNAR	LEFT	multiva3	0.15				1.00	1.00				0.50			0.50
MNAR	LEFT		0.2							1.00				1.00	1.00
MNAR	MID		0.1				0.50	1.00	1.00				0.50		
MNAR	MID	univa	0.15	1.00	1.00				1.00						
MNAR	MID		0.2	1.00				1.00	1.00						
MNAR	MID		0.1				1.00	1.00							1.00
MNAR	MID	multiva2	0.15				1.00	1.00	1.00						
MNAR	MID		0.2				1.00	1.00							1.00
MNAR	MID		0.1				1.00				1.00	1.00			
MNAR	MID	multiva3	0.15							1.00	1.00	1.00			
MNAR	MID		0.2							1.00	1.00	1.00			
MNAR	RIGHT		0.1			0.50	0.50	1.00		0.50		0.50			
MNAR	RIGHT	univa	0.15							1.00	1.00	1.00			
MNAR	RIGHT		0.2			1.00	1.00							1.00	
MNAR	RIGHT		0.1								1.00		1.00		1.00
MNAR	RIGHT	multiva2	0.15					1.00	1.00	1.00					
MNAR	RIGHT		0.2				1.00	1.00	1.00						
MNAR	RIGHT		0.1							0.50	1.00	1.00	0.50		
MNAR	RIGHT	multiva3	0.15							1.00	1.00	1.00			
MNAR	RIGHT		0.2					1.00		1.00					1.00

TABLE XIII. Score Obtained With Respect to the Arithmetic Average Metric (Own Elaboration)

Characteristics of Amputated Datasets				Imputation Method			
				Media	k-NN	k-Means	Hot-Deck
Mechanism	**Type**	**Pattern**	**MR**	$P(\Delta Q)$	$P(\Delta Q)$	$P(\Delta Q)$	$P(\Delta Q)$
MCAR		univa	0.1		**1.00**		
MCAR			0.15			1.00	
MCAR			0.2		1.00		
MCAR		multiva2	0.1			1.00	
MCAR			0.15				1.00
MCAR			0.2		1.00		
MCAR		multiva3	0.1		1.00		
MCAR			0.15				1.00
MCAR			0.2	1.00			
MAR	LEFT	univa	0.1	1.00			
MAR	LEFT		0.15	1.00			
MAR	LEFT		0.2	1.00			
MAR	LEFT	multiva2	0.1		1.00		
MAR	LEFT		0.15		1.00		
MAR	LEFT		0.2		1.00		
MAR	LEFT	multiva3	0.1		1.00		
MAR	LEFT		0.15			1.00	
MAR	LEFT		0.2			1.00	
MAR	MID	univa	0.1		1.00		
MAR	MID		0.15	1.00			
MAR	MID		0.2		1.00		
MAR	MID	multiva2	0.1		1.00		
MAR	MID		0.15		1.00		
MAR	MID		0.2		1.00		
MAR	MID	multiva3	0.1			1.00	
MAR	MID		0.15			1.00	
MAR	MID		0.2			1.00	
MAR	RIGHT	univa	0.1		1.00		
MAR	RIGHT		0.15		1.00		
MAR	RIGHT		0.2	1.00			
MAR	RIGHT	multiva2	0.1			1.00	
MAR	RIGHT		0.15		1.00		
MAR	RIGHT		0.2		1.00		
MAR	RIGHT	multiva3	0.1			1.00	
MAR	RIGHT		0.15			1.00	
MAR	RIGHT		0.2			1.00	
MNAR	LEFT	univa	0.1	1.00			
MNAR	LEFT		0.15		1.00		
MNAR	LEFT		0.2		1.00		
MNAR	LEFT	multiva2	0.1		1.00		
MNAR	LEFT		0.15		1.00		
MNAR	LEFT		0.2		1.00		
MNAR	LEFT	multiva3	0.1		1.00		
MNAR	LEFT		0.15		1.00		
MNAR	LEFT		0.2				1.00
MNAR	MID	univa	0.1		1.00		
MNAR	MID		0.15	1.00			
MNAR	MID		0.2		1.00		
MNAR	MID	multiva2	0.1		1.00		
MNAR	MID		0.15		1.00		
MNAR	MID		0.2		1.00		
MNAR	MID	multiva3	0.1			1.00	
MNAR	MID		0.15			1.00	
MNAR	MID		0.2			1.00	
MNAR	RIGHT	univa	0.1		1.00		
MNAR	RIGHT		0.15			1.00	
MNAR	RIGHT		0.2		1.00		
MNAR	RIGHT	multiva2	0.1				1.00
MNAR	RIGHT		0.15		1.00		
MNAR	RIGHT		0.2		1.00		
MNAR	RIGHT	multiva3	0.1			1.00	
MNAR	RIGHT		0.15			1.00	
MNAR	RIGHT		0.2		1.00		

TABLE XIV. Scores Obtained by IM for Each Metric (Own Elaboration)

Imputation Method	Score obtainedfor each metric				
	$o_1(\Delta Cal)$	$o_2(\Delta Prec)$	$o_3(\Delta Clas)$	$O(\Delta Q)$	G
Media	12.83	12.00	10.00	8.00	42.83
k-NN	23.83	34.50	20.00	34.00	112.33
k-Means	21.33	12.50	16.00	17.00	66.83
Hot-Deck	5.00	4.00	17.00	4.00	30.00

The values in Table XIV are plotted in Fig. 5.

By sorting the values in Table XIV in descending order, the imputation methods for each metric were obtained, according to their order of goodness of fit to impute the set/group of data sets (files).

Regarding the values of the Δ*Cal* metric, the k-NN, k-Means, Mean and Hot-Deck methods, according to their order of goodness of fit, were better. Similarly, considering the values of the Δ*Prec* metric, the k-NN, k-Means, Mean and Hot-Deck methods, according to their order of goodness of fit, were obtained. However, considering the values of the Δ*Clas* metric, the k-NN, Hot-Deck, k-Means and Mean methods resulted. Finally, as for the values of the arithmetic average metric Δ*Q*, the k-NN, k-Means, Mean and Hot-Deck methods resulted according to their order of goodness.

Finally, considering the values of the overall score metric *G*, the k-NN, k-Means, Mean and Hot-Deck methods were ranked according to their order of goodness of fit.

Summarizing, the best imputation methods globally considered turned out to be k-Means and k-NN according to criterion 1, k-NN and k-Means according to criterion 2 of this proposal, and k-Means and k-NN according to the calculation methodology based on the square root of the mean square error shown in [11].

V. Conclusions

This paper has presented an innovative methodology to evaluate the performance of imputation methods, based on metrics derived from data mining processes, instead of the generally used methods based on the root mean square error and its derivatives.

The proposed methodology is applicable to data sets to which data mining processes (e.g. regressions) can be applied, which will provide the information with which the different metrics will be calculated.

The working environment implemented to perform the amputation and subsequent imputation experiments described in [11] was appropriate. It has facilitated the management of the respective original, amputated and imputed files, to which the data mining processes performed with ISW V.9.7 software was applied.

The proposed methodology and the metrics presented have made it possible to arrive at an overall value (since it takes into account all the variables that were amputated and then imputed by various methods), indicative of the performance of each imputation method, expressed in comparable values (since it is based on normalized values of data mining metrics), integrating the results of a multitude of tests representative of different scenarios, with different percentages, diversity of patterns, considering also the three most frequent mechanisms of occurrence of missing data.

The results obtained with the proposed methodology in its different variants of metrics (differences in absolute values and scores) are slightly different. However, they concur that the best imputation methods globally considered are k-NN and K-Means, which also coincides with the global results obtained by the metrics indicated in [11].

The proposed methodology, by contemplating several metrics derived from the DMPs, allows working with only one of them or with all of them simultaneously, to determine the best imputation methods for a given scenario. Moreover, it can be applied to the evaluation of any imputation method, since it works with the imputed files and not with the methods themselves.

This methodology makes it possible to use the DMM generated to evaluate the imputation methods, to perform *a posteriori* predictive data mining process, which constitutes an added value of this proposal.

A. Future Lines of Work

To extend the scope of the proposed methodology, we plan to develop new metrics and indicators. We will use combined algorithms based on mean square error and data mining algorithms applied on the complete files, and then on the files imputed by different methods after having been amputated by different mechanisms.

Acknowledgment

This work has been developed in the context of the Research Project code SIUTIRE0005231TC, of the Resistencia Regional Faculty of the National Technological University, Argentine. We would like to thank the Co-Director of this project, Dr. Marcelo Karanik, for reviewing this work, Dr. Jorge Emilio Monzón, for reviewing the English version, and the scholarship holder, student Alejandro Nadal, for his effort and dedication to the multiple data mining processes.

References

[1] P. Schmitt, J. Mandel, and M. Guedj, "A comparison of six methods for missing data imputation," *Journal of Biometrics & Biostatistics*, vol. 6, no. 1, pp. 1–6, 2015, doi: 10.4172/2155-6180.1000224.

[2] A. Farhangfar, L. A. Kurgan, and W. Pedrycz, "A novel framework for imputation of missing values in databases," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 37, no. 5, pp. 692–709, 2007.

[3] T. Aljuaid and S. Sasi, "Proper imputation techniques for missing values in data sets," in *Proc. 2016 International Conference on Data Science and Engineering (ICDSE)*, 2016.

[4] M. S. Santos, R. C. Pereira, A. F. Costa, J. P. Soares, J. Santos, and P. H. Abreu, "Generating synthetic missing data: A review by missing mechanism," *IEEE Access*, vol. 7, pp. 11651–11667, 2019, doi: 10.1109/access.2019.2891360.

[5] Y. Liu and V. Gopalakrishnan, "An overview and evaluation of recent machine learning imputation methods using cardiac imaging data," *Data*, vol. 2, no. 8, pp. 1–15, 2017, doi: 10.3390/data2010008.

[6] P. J. García-Laencina, J. L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: A review," *Neural Computing and Applications*, vol. 19, no. 2, pp. 263–282, 2010.

[7] M. M. Rahman and D. N. Davis, "Machine learning-based missing value imputation method for clinical datasets," in *IAENG Transactions on Engineering Technologies, Lecture Notes in Electrical Engineering*, vol. 229, pp. 245–257, 2013.

[8] J. M. Jerez, I. Molina, E. A. García-Laencina, N. Ribelles, M. Martín, and L. Franco, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artificial Intelligence in Medicine*, vol. 50, pp. 105–115, 2010.

[9] N. Z. Abidin, A. R. Ismail, and N. A. Emran, "Performance analysis of machine learning algorithms for missing value imputation," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 9, no. 6, pp. 442–447, 2018.

[10] J. Luengo, S. García, and F. Herrera, "On the choice of the best imputation methods for missing values considering three groups of classification methods," *Knowledge and Information Systems*, vol. 32, no. 1, pp. 77–108, 2012.

[11] C. R. Primorac, D. L. La Red Martínez, and M. E. Giovannini, "Metodología de evaluación del desempeño de métodos de imputación mediante una métrica tradicional complementada con un nuevo indicador," *European*

Scientific Journal (ESJ), vol. 16, no. 18, pp. 61–92, 2020.

[12] C. Ballard, J. Rollins, J. Ramos, A. Perkins, R. Hale, A. Dorneich, E. C. Milner, and J. Chodagam, *Dynamic warehousing: Data mining made easy*, IBM Corporation, 2007.

[13] G. Madhu and T. V. Rajinikanth, "A novel index measure imputation algorithm for missing data values: A machine learning approach," in *2012 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 2012, doi: 10.1109/ICCIC.2012.6510198.

[14] D. L. La Red Martínez, M. Karanik, M. Giovannini, M. E. Báez, and J. Torre, "Descubrimiento de perfiles de rendimiento estudiantil: un modelo de integración de datos académicos y socioeconómicos," *Revista Científica Iberoamericana de Tecnología Educativa - Scientific Journal of Educational Technology*, vol. 5, no. 2, pp. 70–83, 2016.

[15] J. Han, M. Kamber, and J. Pei, *Data mining: Concepts and techniques*, 3rd ed., Amsterdam, Netherlands: Elsevier, 2012.

[16] R. J. Roiger, *Data mining: A tutorial-based primer*, 2nd ed., Boca Raton, FL, USA: CRC Press, Taylor & Francis Group, 2016.

[17] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, no. 3, pp. 37–54, 1996.

[18] I. Kononenko and M. Kukar, *Machine learning and data mining: Introduction to principles and algorithms*, Amsterdam, Netherlands: Elsevier, 2007.

[19] S. Chakrabarti, E. Cox, E. Frank, R. H. Güting, J. Han, X. Jiang, M. Kamber, S. S. Lightstone, T. P. Nadeau, R. E. Neapolitan, D. Pyle, M. Refaat, M. Schneider, T. J. Teorey, and I. H. Witten, *Data mining: Know it all*, Amsterdam, Netherlands: Elsevier, 2009.

[20] C. Ballard, N. Harris, A. Lawrence, M. Lowry, A. Perkins, and S. Voruganti, *InfoSphere Warehouse: A robust infrastructure for business intelligence*, IBM Corporation, 2010.

[21] D. L. La Red Martínez and J. C. Acosta, "Aggregation operators review - mathematical properties and behavioral measures," *International Journal of Intelligent Systems and Applications (IJISA)*, vol. 7, no. 10, pp. 63–76, 2015.

[22] P. Chan Chiu, A. Selamat, O. Krejcar, K. Kuok Kuok, E. Herrera-Viedma, and G. Fenza, "Imputation of rainfall data using the sine cosine function fitting neural network," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 7, pp. 39–48, 2021.

David L. la Red Martínez

David L. la Red Martínez obtained a master's degree in computer science and informatics at the National University of the Northeast - UNNE (Argentina) in 2001 and a PhD in systems engineering and computer science at the University of Malaga - UMA (Spain) in 2011. He is currently a full professor at the National University of the Northeast and at the National Technological University - UTN and director of the "Operating Systems and ICT" Research Group at UNNE. For more than 20 years, he has worked in research projects both at national and international level. His research has focused on distributed systems, decision support systems, data communications, ICT in education and educational and health data mining.

Carlos R. Primorac

Carlos R. Primorac obtained a degree in computer science at the National University of the Northeast - UNNE (Argentina) in 2015. He is currently a professor at the National University of the Northeast and a member of the "Operating Systems and ICT" Research Group at UNNE. For more than 6 years, he has worked in research projects at national level. His research has focused on distributed systems, data communications and data imputation.

Traffic Optimization Through Waiting Prediction and Evolutive Algorithms

Francisco García[1], Helena Hernández[2], María N. Moreno-García[2], Juan F. De Paz[1]*, Vivian F. López[2], Javier Bajo[3]

[1] Expert Systems and Applications Lab. University of Salamanca. Plaza de los Caídos s/n. Salamanca (Spain)
[2] Data Mining Research Group. University of Salamanca Plaza de los Caídos s/n. Salamanca (Spain)
[3] Department of Artificial Intelligence, ETSI Informáticos, Universidad Politécnica de Madrid, 28660 Madrid (Spain)

* Corresponding author: fcofds@usal.es

Received 10 May 2022 | Accepted 22 September 2023 | Published 5 December 2023

Abstract

Traffic optimization systems require optimization procedures to optimize traffic light timing settings in order to improve pedestrian and vehicle mobility. Traffic simulators allow obtaining accurate estimates of traffic behavior by applying different timing configurations, but require considerable computational time to perform validation tests. For this reason, this project proposes the development of traffic optimizations based on the estimation of vehicle waiting times through the use of different prediction techniques and the use of this estimation to subsequently apply evolutionary algorithms that allow the optimizations to be carried out. The combination of these two techniques leads to a considerable reduction in calculation time, which makes it possible to apply this system at runtime. The tests have been carried out on a real traffic junction on which different traffic volumes have been applied to analyze the performance of the system.

Keywords

Evolutionary Algorithms, Prediction Systems. Traffic Optimization, Traffic Simulator.

DOI: 10.9781/ijimai.2023.12.001

I. Introduction

According to United Nations data, in 2018 55% of the population was living in urban spaces, the distribution of the urban population varies considerably by region: Northern America 82%, Latin America and the Caribbean 81%, Europe 74%, Oceania 68%, Asia 60% and Africa 43%. The urban population is continuously increasing; it is estimated that 66% of the population will live in urban areas by 2050, an increase of 16% compared to 2008 [1]. These data are very similar to those provided by the United Nations organization since in 2018 it estimated that 68% of the population will live in urban areas in 2050. This increase implies greater traffic congestion in cities due to both the increase in traffic and the unsuitable infrastructures [1]. For this reason, programs such as Horizonte Europa have analyzed global challenges such as climate, energy and mobility, and in particular, intelligent mobility through the optimization of infrastructures. Due to this increase in population and the need to improve infrastructure management, there is a demand to create systems capable of improving traffic efficiency, which will be applied in this project.

Traditional operational research incorporates the use of queuing theory to make predictions about different parameters such as waiting times [2]. The queuing theory approach in which there are usually M/G/s models [3] where M refers to the arrival of vehicles which is represented by a poisson, G the service rate which in certain cases can be modeled by an exponential and finally, s represents the number of servers. From these definitions, it is possible to determine parameters such as waiting times, which will be the object of study of this project. However, classical queuing theory would not take into account parameters that need to be considered, such as the time lost from the moment a traffic light turns green until the cars start moving. For these reasons, simulators such as SUMO [4] are currently being used for time estimation. The SUMO simulator uses an extension of the Gipps model [5] in which aspects such as user reaction time, braking time, or speed differences between the vehicles in the queue are taken into account. However, for this study the aim is not to apply the use of certain equations, but rather to create a system that is capable of estimating waiting times from traffic data obtained from SUMO simulations in order to subsequently perform optimizations.

In order to improve traffic efficiency, studies are mainly based on the analysis of intersections with or without traffic lights [6]. In the study [6] the convenience of introducing traffic lights at an intersection is analyzed by converting a nonlinear integer programming problem to linear integer programming in order to achieve an efficient resolution. The intersection problem is not restricted to decide only when it is more appropriate to introduce a traffic light, but it also involves the problem of dynamically controlling the timing of traffic

Please cite this article in press as:
F. García, H. Hernández, M. N. Moreno-García, J. F. De Paz, V. F. López, J. Bajo. Traffic Optimization Through Waiting Prediction and Evolutive Algorithms, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 96-103, 2025, http://dx.doi.org/10.9781/ijimai.2023.12.001

lights to reduce waiting times [7] through the application of different techniques such as Bayesian networks [8], evolutionary techniques [9], reinforcement learning [10], fuzzy logic [11], [12]... Waiting times are usually associated with vehicles, but it is also relevant to consider pedestrian waiting times since they also have a relevant impact on vehicle waiting times at intersections.

In this project it is proposed a system that allows to cover two aspects, first, the system allows to make an estimation of waiting times through the use of different prediction techniques, which allows to calculate these waiting times without the need of testing with a simulator, which would require a high computational time. On the other hand, the system allows the optimization of traffic light configurations, thus reducing waiting times through evolutionary algorithms. The use of estimators allows a considerable time reduction, which makes this technique more dynamically applicable to traffic changes. The system has been tested on a real intersection on which different traffic flows have been applied in order to analyze the performance of the prediction systems and also of the optimization method applied.

The article is structured as follows: section 2 contains a description of the state of the art, section 3 the proposal for the data analysis and finally sections 4 and 5 the case study and the results obtained.

II. Related Works

Systems for the improvement of mobility in infrastructures are usually based on the management of intelligent traffic lights in which the timing of the different states can be changed dynamically [9]. In this review we will analyze different studies that determine the timing of the different traffic light states in order to reduce the waiting times of vehicles.

Among the studies that can be found are those based on fuzzy logic, which have been carried out for quite some time. For example, there is the work [13] from 1977, in which a study of time intervals and vehicle flow to manage an intersection was carried out. This work includes a model for traffic simulation, in which they consider different aspects in each traffic light cycle such as number of waiting vehicles, queues, saturation, and car delays in order to calculate the optimal time of the traffic lights. Subsequently, these studies were extended to more intersections [14], [15], [7], [16] and the simulator initially defined in [13] was also adapted by incorporating more intersections in one or more roads [17]. In more recent studies [11] a combination of Fuzzy Logic Controllers and genetic algorithms is performed to optimize the management of several intersections with traffic lights, this procedure allows using Fuzzy Logic to establish times, specifically the number of vehicles in the intersection is taken into account and applying fuzzy logic and the Mandami method the time interval of each traffic light is established. Genetic algorithms are used to maximize the number of vehicles crossing the intersections and fuzzy logic for the estimation of the green intervals of the traffic lights. In the paper [7] it is possible to find an extensive study on different works in which different defuzzification and memberships functions are applied. In some works such as [18] the combination of fuzzy logic and a neural network is analyzed to control the delay of the green state of traffic lights taking into account the size of the queue of cars. There are also works that attempt to improve traffic flow from route prediction through the use of regression methods for time estimation and fuzzy logic for the selection of the best route [19].

For the estimation of the time of traffic lights, procedures can be applied in order to determine the congestion levels, thus, in k [20] the congestion level is estimated through a time series from the use of decision trees, regression and neural networks to try to reduce pollution and energy consumption collecting data for five days. In the work [21] a prediction of congestion is also made through the use of neural networks such as LSTM (Long Short-Term Memory) and also regressors such as Support Vector Regression, Random Forest, Gradient Boosting Regression and other statistical techniques and it is verified how these systems are able to predict congestion from the creation of matrices that represent congestion, and to do so they use information from historical data of speeds, road maps, distances. In [22] an estimation of the daily traffic in England and Wales is made from the application of cluster and regression techniques such as Support Vector Regression (SVR) and Random Forest (RF). Likewise in the work [23] the SVR is applied to make a traffic flow prediction, but in this case other aspects such as meteorological factors are considered. In other study [24], traffic flow prediction is carried out through the use of an ARIMA model and an LSTM network that predicts the number of vehicles in 15-minute periods. First, the linear regression feature of the traffic data is captured by using ARIMA, then back-propagation is applied to train the LSTM network and capture the nonlinear features of the data, and finally, both results are combined based on the dynamic weighting of the sliding window. Using three sets of highway data, this method was compared with the other techniques separately (ARIMA, LSTM and EW) and it was determined that the proposed combined model has better prediction effects.

On the other hand, another parameter to be used to describe the traffic flow can be the average speed of cars within a given period of time [25], generally focused on the short term. In this work, recurrent neural networks are explored using historical time data, as well as a number of contextual factors, including additional information such as date, week, etc., to determine how accurate the speed prediction is. A multi-layered RNN (two versions, one with LSTM and one with GRU) is used to learn the sequential traffic data, and a sparse autoencoder is used for the contextual data. Both outputs are merged and delivered to the predictor (neural network) to learn traffic patterns and predict future speed. The model was tested with two real-world data sets and compared with ten frequently used models, k nearest neighbor (k-NN), support vector machine (SVM), decision tree (DT), gradient booting decision tree (GBDT), random forest (RF), stacked autoencoder (SAE), LSTM, GRU, Con-vLSTM, BiLSTM, showing that the proposed model (specifically the version with LSTM) performs better than the rest in terms of stability and accuracy.

Finally, due to the impossibility of considering, with existing algorithms, nonlinear historical data and other uncertain factors that influence peak-hour congestion, hybrid neural network algorithms such as CNN (Convolutional Neural Network) and LSTM are also proposed for short-term prediction of traffic flows based on multivariate analysis [26]. Traffic information is obtained from a Pavement Management System (PMS) that stores data from multiple detectors located throughout California, and weather information (such as temperature, humidity, etc.) from Mesowest. Experimental results show that the combination of CNN and LSTM obtains a high degree of accuracy compared to other models.

Another type of methods used for traffic optimization are reinforcement learning methods. These methods allow an agent to interact in a smart way with the environment in real time. At each instant of time, the agent perceives the environment, evaluates the policy, and performs the optimal action according to the policy. For each action performed by the agent, a reward is assigned according to whether this action brings the agent closer to or further away from the objectives. From previous observation-action pairs and their associated rewards, the agent is able to optimize its policy to maximize the rewards obtained.

Within these reinforcement learning methods, the most commonly used in traffic optimization are Q-learning based methods. The most common is traditional Q-learning with works such as [27], [28], [29], [30] and [31]. Other variants such as Deep Q-learning with works such as [31], [32], and [33]; and Double Deep Q-learning [34] also appear frequently in the state of the art. However, although they are the most common, Q-learning based techniques are not the best performers. The best results are obtained by other algorithms such as SARSA [27] or variants of Actor-Critic, for instance, Traditional Actor-Critic [27], Advantage-Actor-Critic [35] o Deep Deterministic Policy Gradient [36].

In addition to the techniques used for traffic optimization, it is also relevant to consider the alternatives in traffic simulators. There are several software packages for traffic simulation that offer different functions and, therefore, it is common to find articles that use different options according to the needs of the study or, sometimes, several simulators within the same study in order to make comparisons.

For instance, the VISSIM software package can be used to train an algorithm using reinforcement learning to optimize the safety of signalized intersections [37] This same traffic simulator is also used simultaneously with TransModeler [38] as a comparison to demonstrate that the proposed SSAT model offers better performance when applied to simulate mixed traffic on two-way, two-lane roads. Furthermore, in the work [39] the ability of the CORSIM software package to replicate the highway failure process is assessed and a sensitivity analysis is performed on different driver behavior parameters to determine the effect of these on such failures. However, the simulator finally chosen to carry out the tests on the study intersection was SUMO due to the widespread use of this simulator in the scientific field such as [30], where this software is used to obtain traffic information which will be used in a Q-learning algorithm to create a TSC system that maximizes the number of vehicles passing through an intersection; or [9], where it is used to evaluate in real time the performance of the proposed algorithms (swarm heuristic optimization algorithms, PSO) using real-world data (intersection in Turkey) to optimize traffic light control.

As illustrated above, there are studies that make use of regression techniques in traffic analysis, but these studies are focused on traffic flow estimation. In this work, the use of these techniques will be focused on the prediction of waiting times in order to reduce the time in the simulations to determine the behavior of different configurations without the need to perform a simulation. The use of this procedure would allow the system to adapt to different behaviors of the environment without being tied to any simulator, although SUMO will be used for testing. Subsequently, this data can be used with different optimization techniques, which will reduce computation time and thus improve its applicability to dynamic environments that require constant traffic adaptations.

III. Proposal

The proposed system consists of three components as shown in Fig. 1, the first of which would be the waiting time prediction part that allows estimating waiting times based on the time intervals of the traffic lights. In addition, it must be considered that it is necessary for the traffic lights to comply with some temporal relationships that are defined as contracts in such a way that the time intervals of a traffic light affect the time intervals of the rest. The second component is optimization. Optimization uses the first estimation component to generate from evolutionary algorithms an optimal configuration of traffic light times to reduce a certain parameter, in this case the waiting time of vehicles. The last component is the simulator, which is initially used to generate waiting time data from different configurations and use this information for the first estimation component; subsequently, the simulator component is replaced by the predictor component when performing the optimizations in the optimization component.

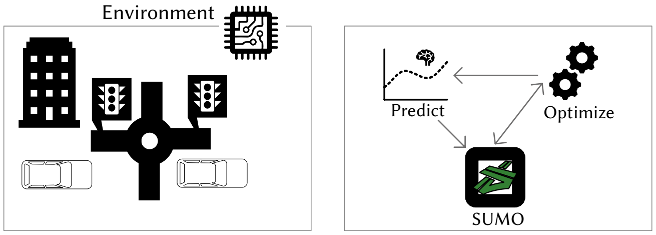


Fig. 1. System Components.

A. Prediction Component

The prediction component acts as a substitute for the traffic simulator during the optimization process. Therefore, its function is to estimate the waiting time at the traffic lights from the time intervals provided to the traffic lights. In this prediction component, techniques based on regressors and neural networks have been incorporated to estimate the waiting times. Specifically, Random-Forest [40], AdaBoost based on a decision tree [41], Bagging also based on a decision tree [42], ExtraTrees using the Gini index for the gain [43], and deep learning techniques and neural networks [44] have also been included to make the predictions within the prediction model.

Four different architectures are used within neural networks: a neural network with a single hidden layer, a neural network with multiple layers (specifically, 14 layers in the best performing one), a neural network with multiple layers and jump connections (specifically, 16 layers in the best performing one), and, finally, an LSTM.

All these models are trained using data generated by the simulator before the optimization process, but future work could use data obtained in real environments and remove the simulator completely from the system. In addition, adding inputs related to road conditions and structure to the predictors could move towards real-time traffic optimization by adapting to actual flow conditions.

B. Optimization

Traditionally, in traffic estimation studies, the optimizer launches multiple simulations with different parameters in the traffic simulator in order to evaluate its efficiency. These simulations are complex and have a high time cost, resulting in an inefficient optimization process. In this work, the optimizer does not communicate at any time with the traffic simulator; instead, the optimizer communicates only with the prediction component, greatly accelerating the optimization process in exchange for a small penalty in the time cost.

This optimization is performed using a particle optimization algorithm, but considering that, in this case, the particles correspond to configurations of the traffic lights and , therefore, there are some relationships and restrictions between them that must be fulfilled as their position is updated. For this reason, in Fig. 2 the information of the restrictions and relations between the traffic lights is included in order to limit the value in each of the iterations and thus obtain valid solutions. Each particle at time instant t is represented by $x_i(t)$, and will contain as many values as variables are being optimized. X is the set of particles, $v_i(t)$ is the velocity with which particle i moves, c_1 is the cognitive acceleration factor, c_2 is the social acceleration factor, p_i is the most optimal solution calculated for particle i, p stores the set of most optimal values for all particles, p_{beast} is the best calculated global solution. C_i contains the constraint for traffic light i, c_i^l is the lower bound for constraint i, and c_i^u is the upper bound for constraint i, C contains the set of constraints for all traffic lights, T is a set of values of the estimation of the time lost for each particle i.

```
// Update particles
  repeat
    foreach $(x_i) \in X$ do
      $v_i(t+1) = v_i(t) + c_1 \cdot rand \cdot (p_i - x_i(t)) + c_1 \cdot rand \cdot (p_{best} - x_i(t))$
      $x_i(t+1) = x_i(t) + v_i(t+1)$
    end
    // Update constraints according to the relation among semaphores
    foreach $c_i \in C$ do
      $c_i$ = updateConstraint (X, C)
    end
    // Update particles according to the constraints
    repeat
      foreach $(x_i) \notin c_i$ do
        if $x_i <$ then
          $x_i(t+1) = c_i^l$
        else
          $x_i(t+1) = c_i^u$
        end
      end
    until $\forall ix_i \in c_i$;
    // Predict time loss with regressor
    T = timeLossParticle (X, regressorCars, C, listSemaphores )
    // Update local and global best
    (p, $p_{beast}$) = updateLocalGlobal (T)
  until;
// X matrix with particples
// regressorCars regressor to predict loos time
// C matrix with constraints semaphores
// listSemaphores list with each semaphore
```

Fig. 2. Optimization Process.

C. Traffic Simulator

In this work, after evaluating several alternatives, SUMO was used as a traffic simulator. SUMO is a spatially continuous microscopic traffic simulator. This means that SUMO simulates vehicle-to-vehicle traffic flow in a non-discretized space.

SUMO uses the traffic model proposed by Stefan Krauß [45], [46]. This model is intended to be a simpler alternative to previous proposals, but, at the same time, to accurately capture traffic dynamics. The model is based on car tracking, i.e., the behavior of one car is conditioned by the positions and speeds of neighboring cars. Specifically, the S. Krauß model is based on the calculation of a maximum speed at which a car can go in such a way that it is impossible for it to collide with the cars it is following, considering a specific deceleration capacity and reaction time. Cars try to go at the maximum safe speed at all times. In addition, it introduces a stochastic term in the calculation of the current speed as a function of acceleration, which introduces a random element into the simulation.

Ultimately, continuous models offer greater accuracy at the cost of a performance penalty. Similarly, microscopic simulations offer greater accuracy than macroscopic simulations, which simulate the behavior of cars at a higher accuracy, again at the cost of a performance penalty. The model proposed by Krauß, despite being one of the simplest among the spatially continuous microscopic models, also carries a high time complexity that scales linearly with respect to the number of cars and the number of time instants simulated.

IV. Case Study

In order to simulate the intersection, the SUMO tool was used, which has a series of console commands that allow generating the flow of cars and pedestrians to carry out the simulation. In order to facilitate the continuous use of these commands and to automatize the process, all of them were grouped in a Python script that makes the necessary calls and creates the files required by SUMO to run the simulation. In the following section, we will describe how both the generation of cars and pedestrians from this file and the creation of the traffic light logic work.

For the generation of cars, we decided to create several flows for each of the routes that make up the intersection shown in Fig. 3, each of which can have a different number of cars, which is indicated as a variable within the Python script. This was done to have more control over the vehicles and to be able to make a model which was closer to reality, since otherwise the random generation could put all the cars on the same route.

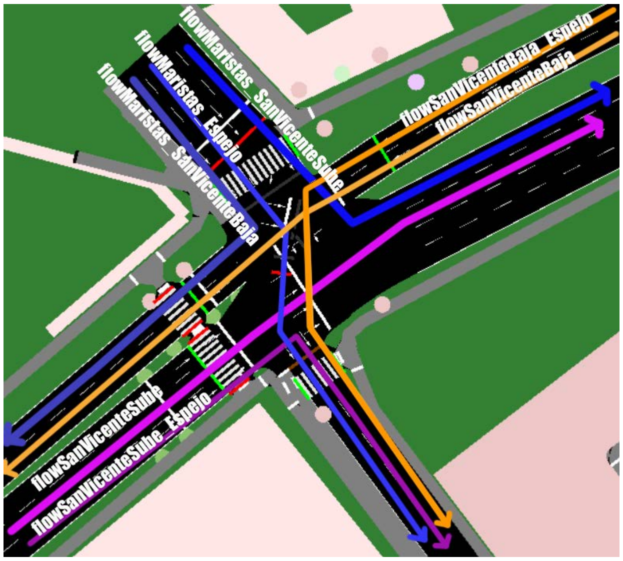
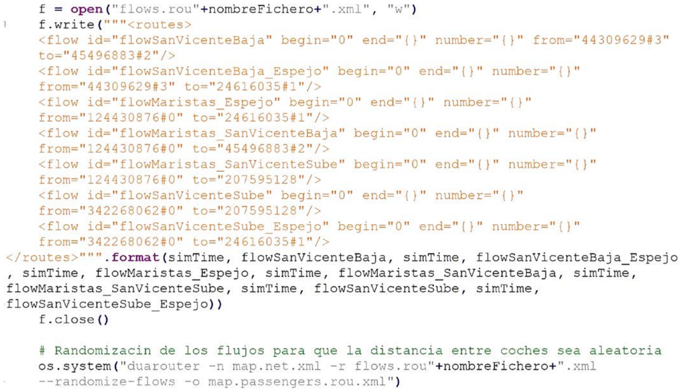

flowMaristas_SanVicenteSube
flowMaristas_Espejo
flowMaristas_SanVicenteBaja
flowSanVicenteBaja_Espejo
flowSanVicenteBaja
flowSanVicenteSube
flowSanVicenteSube_Espejo

Fig. 3. Car flows of the different routes.



```
    f = open("flows.rou"+nombreFichero+".xml", "w")
    f.write("""<routes>
    <flow id="flowSanVicenteBaja" begin="0" end="{}" number="{}" from="44309629#3"
    to="45496883#2"/>
    <flow id="flowSanVicenteBaja_Espejo" begin="0" end="{}" number="{}"
    from="44309629#3" to="24616035#1"/>
    <flow id="flowMaristas_Espejo" begin="0" end="{}" number="{}"
    from="124430876#0" to="24616035#1"/>
    <flow id="flowMaristas_SanVicenteBaja" begin="0" end="{}" number="{}"
    from="124430876#0" to="45496883#2"/>
    <flow id="flowMaristas_SanVicenteSube" begin="0" end="{}" number="{}"
    from="124430876#0" to="207595128"/>
    <flow id="flowSanVicenteSube" begin="0" end="{}" number="{}"
    from="342268062#0" to="207595128"/>
    <flow id="flowSanVicenteSube_Espejo" begin="0" end="{}" number="{}"
    from="342268062#0" to="24616035#1"/>
</routes>""".format(simTime, flowSanVicenteBaja, simTime, flowSanVicenteBaja_Espejo
, simTime, flowMaristas_Espejo, simTime, flowMaristas_SanVicenteBaja, simTime,
flowMaristas_SanVicenteSube, simTime, flowSanVicenteSube, simTime,
flowSanVicenteSube_Espejo))
    f.close()

    # Randomizacin de los flujos para que la distancia entre coches sea aleatoria
    os.system("duarouter -n map.net.xml -r flows.rou"+nombreFichero+".xml
    --randomize-flows -o map.passengers.rou.xml")
```

Fig. 4. Flow generation in the xml file.

The value of these variables will be written inside an xml file (*flows.rou.xml*, Fig. 4) specifying the above-mentioned routes so that SUMO can understand them, that is, indicating the ids of the origin (*from*) and destination (*to*). In addition, the time that these flows will last (*end*), is also included, which must coincide with the time that the simulation is expected to last, and randomness is added to the frequency with which the cars of a flow are generated, since by default a uniform frequency is used. The following variable values were used for this case study:

- **simTime:** 1000 (Simulation duration in milliseconds)
- **flowSanVicenteBaja:** 20 cars
- **flowSanVicenteBaja_Espejo:** 20 cars
- **flowMaristas_Espejo:** 40 cars
- **flowMaristas_SanVicenteBaja:** 20 cars
- **flowMaristas_SanVicenteSube:** 20 cars

- **flowSanVicenteSube:** 20 cars
- **flowSanVicenteSube_Espejo:** 20 cars

Therefore, a maximum of 160 cars are generated in 1000 ms in total.

The generation of pedestrians is less complex since it is not divided into flows (since it is not possible to specify routes for them), so it is only necessary to indicate the maximum total number of pedestrians in the entire crossing (100 in this case study). In order to carry out this task, it was necessary to use the *randomTrips* tool included with SUMO, a Python script that allows to create a random file with pedestrian trips (*map.pedestrians.trips.xml*) by means of the following command:

```
python "%SUMO_HOME%\\tools\\randomTrips.py" -n map.net.xml -o map.pedestrians.trips.xml -r map.pedestrians.rou.xml -e simTime -p pPed -l --pedestrians --max-distance 500
```

Where simTime is the total simulation time (1000 ms), used to indicate for how many milliseconds pedestrians have to be generated, and pPed is the repetition rate, obtained by dividing the simulation time by the number of pedestrians. This is because the script generates pedestrians with a constant frequency of 1/pPed per second, so if 100 pedestrians must be generated in 1000 ms, the frequency should be 1000/100. Furthermore, the *--max-distance* option was used to set the maximum length of the trips, so that pedestrians would not be circulating for too long.

Finally, for the traffic light logic, an additional file (*traffic_lights.add.xml*) containing the durations of the green, yellow, red and amber phases for each of them. The goal was to reproduce the real operation of the traffic lights at the intersection, but at the same time allow to modify their durations to a certain extent. For this purpose, four variables are used, as shown in Fig. 5, from which the value of the other phases of the traffic lights are calculated so that the real configuration is respected.

- **green1:** Green time of the traffic light of San Vicente Uphill (Must be less than or equal to green2)
- **green2:** Green time of the traffic light of San Vicente Downhill
- **green3_d:** Green time of the traffic light of Maristas (right lanes, must be greater than or equal to amber3_i)
- **yellow3_i:** Amber time of the traffic light of Maristas (left lanes) The green time will be calculated by subtracting this value from green3_d, so that the left lanes are at most the same time on green as the right lanes and, if they last less, the rest will be on amber.

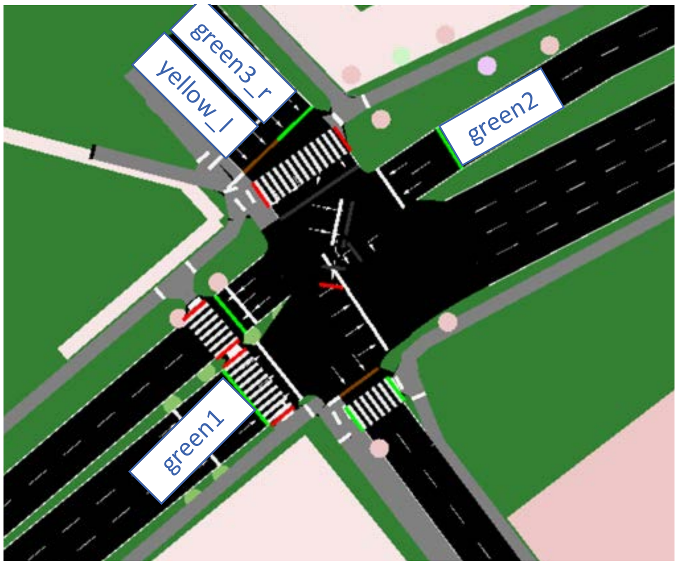


Fig. 5. Variables for traffic lights durations.

V. Results

In this section, the results obtained with the proposed method are discussed, both those of the models for estimating the time lost by vehicles at traffic lights (subsection V.B) and those of the traffic optimization algorithm (subsection V.C). In addition, there is a section in which we discuss why we consider only the time lost by vehicles instead of both vehicles and pedestrians (subsection V.A).

A. Time Lost by Pedestrians and Time Lost by Vehicles

Since the methods consulted in the state of the art only consider the time lost by vehicles to perform traffic optimization, at the beginning of this work, one of the novelties intended to be included was to consider the time lost by pedestrians at traffic lights when performing this optimization. However, when evaluating the results of the optimization using the proposed method, the total lost time (sum of the time lost by vehicles and the time lost by pedestrians) predicted for specific traffic signal times was far from the time calculated by the traffic simulator.

In order to analyze this discrepancy, 1000 runs of the traffic simulator were performed with the same traffic light times, specifically, those predicted as optimal by the optimizer. In these runs, both the time lost by pedestrians and the time lost by vehicles were collected, analyzed and plotted as the density plot shown in Fig. 6. In addition to the optimal times, two other tests were also performed with different values for traffic light times but with similar results. The conclusion of these tests is that introducing randomness in the pedestrian paths introduced a standard deviation of less than 1 second in the time lost by vehicles, but of about 3.5 seconds in the time lost by pedestrians with a difference of more than 20 seconds between the minimum and maximum. In contrast, by repeating these tests with constant pedestrian recoveries and introducing randomization in the vehicle paths, the standard deviation of both lost times is very close to zero.

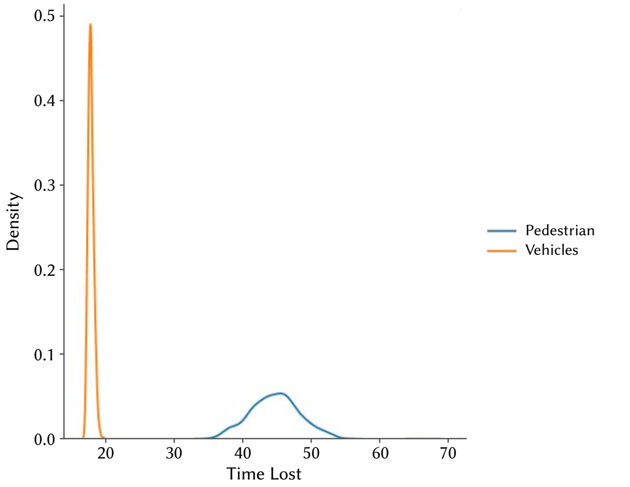


Fig. 6. Density plot of time lost by vehicles and pedestrians for different pedestrian random paths.

The reason behind this disparity in the values of time lost by pedestrians when randomness is introduced for pedestrians, but not when randomness is introduced for vehicles, is due to the implementation of the traffic simulator used, SUMO. Specifically, SUMO allows us to specify the number of vehicles that will make a route between a specific origin and a specific destination, but it generates the pedestrian routes in a completely random way.

Thus, in order to guarantee reproducibility and considering that this discrepancy in the results was due to the implementation of the traffic simulator itself, there are two possible solutions to the problem: eliminating the randomness introduced in the pedestrians or considering only the time lost by vehicles, which have a smaller deviation. Since limiting the randomness to only vehicles could introduce a bias towards a certain number of pedestrian paths, it was decided to consider only the time lost by vehicles to perform the optimization.

B. Comparison Between Models for Estimating Lost Time at Traffic Lights

As mentioned in Section 3, this project has analyzed the use of several artificial intelligence models to estimate the time lost by vehicles at traffic lights. The methodology used to evaluate these models is described below and a comparison of results is provided.

In order to train these estimation models, we used a dataset generated from the lost time at traffic lights retrieved from multiple simulations for different traffic light times using SUMO. Specifically, these simulations were performed on the scenario described in the case study. A total of 625,000 simulations were performed corresponding to all possible combinations giving values between 1 and 50 seconds to each of the green times of the different traffic lights. In addition, in the case of neural network architectures [43], the hyperparameters were selected using the Bayesian hyperparameter as the tuning method and the mean MAE over a cross-validation of 10 folds as the evaluation criteria. The optimizer used was Adam and the batch size (131,072 samples) was selected to maximize GPU utilization.

To evaluate the performance of each of the models, cross-validation of 10 iterations and the mean absolute error (MAE) metric on each of them was used. The mean of the results over these 10 iterations is given in Table I with a 95% confidence interval.

TABLE I. Mean MAE and NMAE of the 10-Fold Cross-Validation of Each Method With 95% Confidence Interval

Method	MAE	NMAE
Random Forest	0.094 ± 0.000	0.00083 ± 0.00000
Ada Boost	5.488 ± 0.234	0.04840 ± 0.00206
Bagging	0.094 ± 0.000	0.00083 ± 0.00000
Extra Trees	0.006 ± 0.000	0.00005 ± 0.00000
LSTM	1.629 ± 0.004	0.01437 ± 0.00004
Shallow Network	1.761 ± 0.009	0.01553 ± 0.00008
Deep Network	1.143 ± 0.041	0.01008 ± 0.00036
Residual Network	0.873 ± 0.007	0.00770 ± 0.00006

Even though it would be possible to determine which methods perform better from the values available in this table, it was decided to use the Mann Whitney hypothesis validation test to ensure that this assessment has a certain statistical reliability. Specifically, two separate tests were performed for all possible pairs of methods. The first one had as null hypothesis the equality of the results between pairs of methods and as alternative hypothesis the inequality of the results between them. The second test had as the null hypothesis the inferiority of the results of the first method and the alternative hypothesis the superiority of the results of the first method. The results of these tests are shown in Fig. 7.

According to these graphs, it can be shown that the results of the Bagging Regressor and the Random Forest Regressor are equivalent, while the rest of the methods are quite different from each other. On the other hand, it can be observed that the results of the Extra Trees Regressor outperform the results of the other methods while the results of the Ada Boost Regressor are inferior to the rest. Furthermore, it can be observed that the results of neural network based methods are worse than those of traditional Machine Learning algorithms, with the exception of Ada Boost. Based on these observations, it was decided to use the Extra Trees Regressor algorithm as a method for estimating the waiting time at traffic lights in order to optimize traffic flow.

C. Traffic Optimization

The traffic optimization experiments were developed using Python scripts which use a particle optimization algorithm implemented in the pyswarm library. Two implementations were performed. The first one used the SUMO simulator to calculate the waiting times of vehicles at traffic lights and the second one used the Extra Trees Regressor algorithm to calculate an approximation. In the first case, the run lasted 1 day, 7 hours, 33 minutes, and 48 seconds. In the second case, the run lasted 2 minutes and 26 seconds. Although the difference in execution time is dramatic, the results obtained are very similar. Specifically, in the first case, the average waiting time for vehicles was 16.66 seconds and in the second case, 17.26 seconds.

VI. Conclusions and Future Enhancements

We have developed a system capable of optimizing traffic based on particle flooding which improves its performance by replacing the traffic simulator with an estimation system based on machine learning algorithms. Several estimation methods have been analyzed and the one with the best results, the Extra Trees Regressor, has been selected. Finally, the loss of precision in the results when using our method has been evaluated and it has been observed that the resulting waiting time when using the approximator is 0.6 seconds longer than when using the simulator. However, the computation time when using the simulator (113627.74892 seconds) is up to 777 times longer than when using the approximator (146.138249 seconds).

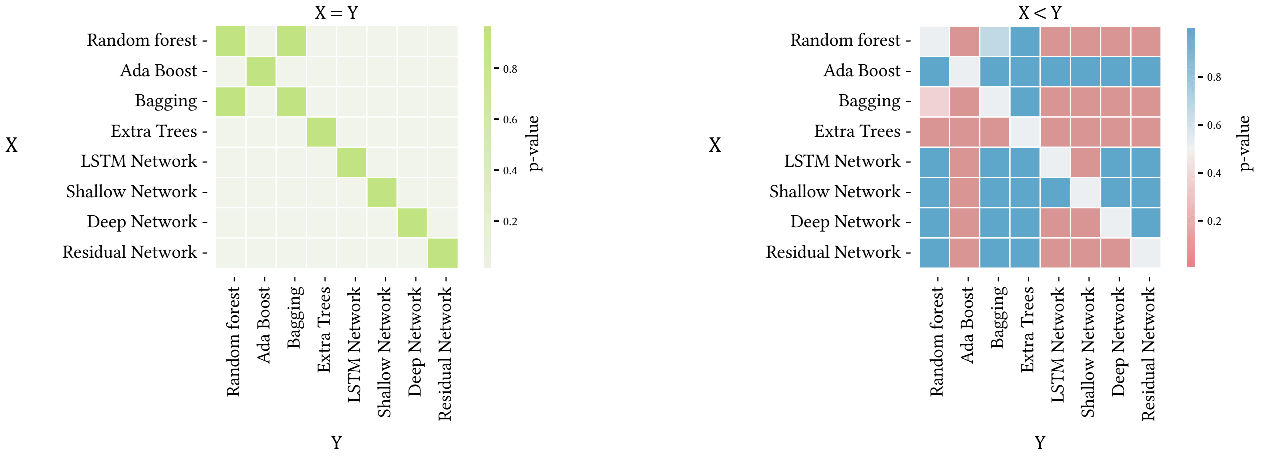


Fig. 7. P-value matrices of the Mann Whitney test for equality and inferiority of the results of the different methods, respectively.

In future studies, we will analyze the use of other traffic optimization algorithms and their compatibility with our Extra Trees Regressor-based approach. Furthermore, mechanisms will be studied to allow the approximation algorithm to be able to generalize to other intersections without the need for a complete retraining. Finally, we will consider the development of a system capable of collecting data to train the estimator automatically by analyzing images obtained by cameras implanted in the traffic lights.

Acknowledgment

This work was supported by the Spanish Agencia Estatal de Investigación. Project Monitoring and tracking systems for the improvement of intelligent mobility and behavior analysis (SiMoMIAC). Sistema de procesamiento distribuido para la optimización en movilidad y transporte - Distributed processing system for mobili-ty and transport optimization PID2019-108883RB-C21 / AEI / 10.13039/501100011033. PID2019-108883RB-C22 - Sistema de monitorizacion y control para el apoyo a la toma de decisiones en el mantenimiento de infraestructuras. Francisco García Encinas' research was partly supported by the Spanish Ministry of Universi-ties (FPU Fellowship under Grant FPU19/02455).

References

[1] S.M. Khan, M. Rahman, A. Apon, M. Chowdhury, "Chapter 1 - Characteristics of Intelligent Transportation Systems and Its Relationship With Data Analytics", *Data Analytics for Intelligent Transportation Systems*, Elsevier, pp. 1-29, 2017.

[2] C. Zato, A. de Luis, J. Bajo, J.F. de Paz, J.M. Corchado, "Dynamic model of distribution and organization of activities in multi-agent systems", *Logic Journal of the IGPL*, vol. 20 no. 3, pp. 570-578, 2012.

[3] D.A. Menasce, "Trade-offs in designing web clusters", *IEEE Internet Computing*, vol. 6, no. 5, pp. 76-80, 2002.

[4] P. Alvarez Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.P. Flötteröd, R. Hilbrich, Leonhard Lücken, J. Rummel, P. Wagner, Evamarie Wießner, "Microscopic Traffic Simulation using SUMO", *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2018.

[5] D. Krajzewicz, G. Hertkorn and P. Wagner, Christian Rössel, "SUMO (Simulation of Urban MObility) An open-source traffic simulation", in *MESM2002 Proceedings*, Comingout Okt., 2002.

[6] Y. Zhang, R. Su, "An optimization model and traffic light control scheme for heterogeneous traffic systems", T*ransportation Research Part C: Emerging Technologies*, vol. 124, 102911, 2021.

[7] C. Karakuzu, O. Demirci, "Fuzzy logic based smart traffic light simulator design and hardware implementation", *Applied Soft Computing*, vol. 10, no.1, pp. 66-73, 2010.

[8] X. Zhengxing, J. Qing, N. Zhe, W. Rujing, Z. Zhengyong, H. He, S. Bingyu, W. Liusan, W. Yuanyuan, "Research on intelligent traffic light control system based on dynamic Bayesian reasoning", *Computers & Electrical Engineering*, vol. 84, 2020.

[9] S.A.Celtek, A. Durdu, M.E.M. Alı, "Real-time traffic signal control with swarm optimization methods", *Measurement*, vol. 166, 108206, 2020.

[10] M. Greguri´c, M. Vuji´c, "Charalampos Alexopoulos and Mladen Mileti. Application of Deep Reinforcement Learning in Traffic Signal Control: An Overview and Impact of Open Traffic Data", *Applied Sciences*, vol. 10, 4011, 2020.

[11] S.M. Odeh, A.M. Mora, M.N. Moreno, J.J.Merelo, "A Hybrid Fuzzy Genetic Algorithm for an Adaptive Traffic Signal System", *Advances in Fuzzy Systems*, 2015.

[12] A. Ikidid, A. El Fazziki, M. Sadgal, "Multi-Agent and Fuzzy Inference-Based Framework for Traffic Light Optimization", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 2, pp. 88-97, 2023.

[13] C. Pappis, E. Mamdani, "A fuzzy logic controller for a traffic junction," *IEEE transactions on Systems, Man and Cybernetics, SMC-7/10*, pp. 707-717, 1977.

[14] M. Nakatsuyama, H. Nagahashi, N. Nishizuka, "Fuzzy logic phase controller for traffic functions in the one-way arterial road", *IFAC 9th Triennial World Congress*, Pergamon Press, pp. 2865-2870, 1984.

[15] J. Favilla, A. Machion, F. Gomide "Fuzzy traffic control: adaptive strategies" *Second IEEE International Conference on Fuzzy Systems II*, pp. 506-511, 1993.

[16] S. Komsiyah, E. Desvania, "Traffic Lights Analysis and Simulation Using Fuzzy Inference System of Mamdani on Three-Signaled Intersections", *Procedia Computer Science*, vol. 179, pp. 268-280, 2021.

[17] C. H. Chou, J. C.Teng, "A fuzzy logic controller for traffic junction signals", *Information Sciences*, vol. 143, no. 1–4, pp. 73-97, 2002.

[18] X. Fan, Y. Liu, "Alterable-Phase Fuzzy Control Based on Neutral Network", *Journal of Transportation Systems Engineering and Information Technology*, vol. 8, no. 1, pp. 80-85, 2008.

[19] A. S. Tomar, M. Singh, G. Sharma, K. V. Arya, "Traffic Management using Logistic Regression with Fuzzy Logic", *Procedia Computer Science*, vol. 132, pp. 451-460, 2018.

[20] T.S. Tamir, G. Xiong, Z. Li, H. Tao, Z. Shen, B. Hu, H.M. Menkir, "Traffic Congestion Prediction using Decision Tree", *Logistic Regression and Neural Networks*, IFAC-PapersOnLine, vol. 53, no. 5, pp. 512-517, 2020.

[21] M. Bai, Y. Lin, M. Ma, P. Wang, L- Duan, PrePCT: "Traffic congestion prediction in smart cities with relative position congestion tensor", *Neurocomputing*, vol. 444, pp. 147-157, 2021.

[22] A. Sfyridis, P. Agnolucci, "Annual average daily traffic estimation in England and Wales: An application of clustering and regression modelling", *Journal of Transport Geography*, vol. 83, 102658, 2020.

[23] X. Bao, D. Jiang, X. Yang, H. Wang, "An improved deep belief network for traffic prediction considering weather factors", *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 413-420, 2021.

[24] S. Lu, Q, Zhang, G. Chen, D. Seng, "A combined method for short-term traffic flow prediction based on recurrent neural network", *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 87-94, 2021.

[25] L. Qu, J. Lyu, W. Li, D. Ma, H. Fan, "Features injected recurrent neural networks for short-term traffic speed prediction", *Neurocomputing*, vol. 451, pp. 290-304, 2021.

[26] S. Narmadha, V. Vijayakumar, "Spatio-Temporal vehicle traffic flow prediction using multivariate CNN and LSTM model", *Materials Today: Proceedings*, 2021.

[27] M. Aslani, S. Seipel, M.S. Mesgari, M. Wiering, "Traffic signal optimization through discrete and continuous reinforcement learning with robustness analysis in downtown tehran", *Advanced Engineering Informatics*, vol. 38, pp. 639–655, 2018.

[28] M. Abdoos, A.L.C. Bazzan, "Hierarchical traffic signal optimization using reinforcement learning and traffic prediction with long short term memory", *Expert Systems with Applications*, vol. 171, pp. 114580, 2021.

[29] M. Essa, T. Sayed, "Self-learning adaptive traffic signal control for real-time safety optimization", *Accident Analysis & Prevention*, vol. 146, pp. 105713, 2020.

[30] H. Joo, S.H. Ahmed, Y. Lim, "Traffic signal control for smart cities using reinforcement learning", *Computer Communications*, vol. 154, pp. 324-330, 2020.

[31] E. Walraven, M.T.J. Spaan, B. Bakker, "Traffic flow optimization: A reinforcement learning approach", *Engineering Applications of Artificial Intelligence*, vol. 52, pp. 203-212, 2016.

[32] T.M. Aljohani, A. Ebrahim, O. Mohammed, "Real-Time metadata-driven routing optimization for electric vehicle energy consumption minimization using deep reinforcement learning and Markov chain model", *Electric Power Systems Research*, vol. 192, pp. 106962, 2021.

[33] S. Koh, B. Zhou, H. Fang, P. Yang, Z. Yang, Q. Yang, L. Guan, Z. Ji, "Real-time deep reinforcement learning based vehicle navigation", *Applied Soft Computing*, vol. 96, pp. 106694, 2020.

[34] Y. Gong, M. Abdel-Aty, J. Yuan, Q. Cai, "Multi-Objective reinforcement learning approach for improving safety at intersections with adaptive traffic signal control", *Accident Analysis & Prevention*, vol. 144, pp. 105655, 2020.

[35] H. Maske, T. Chu, U. Kalabić, "Control of traffic light timing using decentralized deep reinforcement learning", *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 14936-14941, 2020.

[36] Z. Li, H. Yu, G. Zhang, S. Dong, C.Z. Xu, "Network-wide traffic signal control optimzation using a multi-agent deep reinforcement learning",

Transportation Research Part C: Emerging Technologies, vol. 125, pp. 103059, 2021.

[37] M. Essa, T. Sayed, "Self-learning adaptive traffic signal control for real-time safety optimization", Accident Analysis & Prevention, vol. 146, 105713, 2020.

[38] J. Sun, H. Liu, Z. Ma, "Modelling and simulation of highly mixed traffic flow on two-lane two-way urban streets", *Simulation Modelling Practice and Theory*, vol. 95, pp. 16-35, 2019.

[39] A. Kondyli, I. Soria, A. Duret, L. Elefteriadou, "Sensitivity analysis of CORSIM with respect to the process of freeway flow breakdown at bottleneck locations", *Simulation Modelling Practice and Theory*, vol. 22, pp. 197-206, 2012.

[40] L. Wiene, A. Liaw, M. Wiener "Classification and regression by randomForest", *R News*, 2/3, pp. 18-22, 2002.

[41] Y. Freund, R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", *Journal of Computer and System Sciences*, vol. 55, no.1, pp. 119-139, 1977.

[42] D. Hernández-Lobato, G. Martínez-Muñoz, A. Suárez, "Empirical analysis and evaluation of approximate techniques for pruning regression bagging ensembles", *Neurocomputing*, vol. 74, no. 12–13, pp. 2250-2264, 2011.

[43] P. Geurts, D. Ernst, I. Wehenkel, "Extremely randomized trees", *Machine Learning*, vol. 63, no. 1, pp. 3-42, 2006.

[44] F. García Encinas, H. Hernández Payo, J.F. de Paz Santana, M.N. Moreno García, J. Bajo Pérez, "Estimating Time Lost on Semaphores with Deep Learning". *New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence*, vol. 1410, Cham, Springer, 2022. https://doi.org/10.1007/978-3-030-87687-6_4.

[45] S. Krauss, P. Wagner, C. Gawron "Metastable States in a Microscopic Model of Traffic Flow", *Physical Review E*, vol. 55, no. 5, pp. 5597–602, 1997, https://doi.org/10.1103/PhysRevE.55.5597.

[46] S. Krauss, "Microscopic modeling of traffic flow: Investigation of collision free vehicle dynamics", 1998.

Francisco García Encinas

Francisco García Encinas studied a bachelor's degree in Computer Engineering at the University of Salamanca, Spain, from 2015 to 2019. Later, from late 2018 to mid-2019 he worked as a scientific software programmer at ALF-USAL research group at the University of Salamanca. In 2020, he got a master's degree in Intelligent Systems at the same university, and, nowadays, he is a Ph.D. student at ESALab research group.

Helena Hernández Payo

Helena Hernández Payo graduated from the University of Salamanca in 2021 where she obtained a bachelor's degree in Computer Engineering. From 2020 to 2021 she did an internship as software programmer for the research group GRIAL and at the same time contributed to the creation of the simulation described in this article as part of a collaboration scholarship at the University of Salamanca. Currently she is working as a Big Data developer at Viewnext.

María N. Moreno García

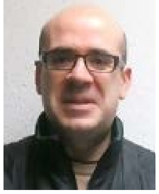
María N. Moreno García is currently Full Professor at the University of Salamanca, Spain, and head of the Data Mining Research Group of the same University (mida.usal.es). She has been a research scholar at the Intelligent System Lab of the University of Bristol, UK, and at the College of Computing and Digital Media of the DePaul University in Chicago, USA. Her research interests are in the areas of Data Science, Machine Learning and their application in different domains.

Juan Francisco De Paz Santana

Juan Francisco De Paz Santana received the degree in technical engineering in systems computer sciences, in 2003, and the Engineering degree in computer sciences, the degree in statistics, and the Ph.D. degree in computer science from the University of Salamanca, Spain, in 2005, 2007, and 2010, respectively. He is currently a Full Professor with the University of Salamanca, where he is also a Researcher with the Expert Systems and Applications Laboratory (ESALab). He has been a coauthor of published articles in several journals, workshops, and symposiums.

Vivian Félix López Batista

Vivian Félix López Batista received a PhD. in Computer Science from the University of Valladolid in 1996. At present she is a Full Professor of Computer Science at the University of Salamanca (Spain) where she has been since 1998. Member of the Data Mining Group (http://mida.usal.es/). She has done research on Natural Language Processing, Machine Learning and Neural Networks. She has also papers published in recognized journals, workshops and symposiums, books, and book chapters in these topics. She performed a research stay in the Center for Computational Science, University of Miami.



Javier Bajo

Javier Bajo received a PhD in computer sciences, in 2003, from the University of Salamanca, Spain. He is currently a Full Professor at the Department of Artificial Intelligence at the Universidad Politécnica de Madrid. He is also Director at the UPM Research Center in Artificial Intelligence. He has been a coauthor of more than 300 articles published in recognized journals, international conferences, workshops, and symposiums.

An Effective Prediction Approach for the Management of Children Victims of Road Accidents

F. Saadi[1]*, B. Atmani[2], F. Henni[3], H.Benfriha[4], Z.Addou[5], R. Guerbouz[6]

[1] Laboratoire d'Informatique d'Oran (LIO), University Oran 1 (Algeria)
[2] Laboratoire d'Informatique d'Oran (LIO), University of Mostaganem (Algeria)
[3] Computer Science and new Technologies Lab (CSTL), University of Mostaganem (Algeria)
[4] Laboratoire d'Informatique d'Oran (LIO),Institut teccart (Canada)
[5] Réanimation Polyvalente Pédiatrique EHS Canastel, University of Oran 1 (Algeria)
[6] Service Neurochirurgie, Etablissement Hospitalo Universitaire Oran (Algeria)

* Corresponding author: saadi_fatima@hotmail.fr

Received 28 October 2022 | Accepted 23 December 2022 | Published 5 February 2024

Abstract

Road traffic generates a considerable number of accidents each year. The management of injuries caused by these accidents is becoming a real public health problem. Faced with this latter, we propose a new clinical decision making approach based on case-based reasoning (CBR) and data mining (DM) techniques to speed up and improve the care of an injured child. The main idea is to preprocess the dataset before using K Nearest Neighbor (KNN) Classification Model. In this paper, an efficient predictive model is developed to predict the admission procedure of a child victim of a traffic accident in pediatric intensive care units. The evaluation of the proposed model is conducted on a real dataset elaborated by the authors and validated by statistical analysis. This novel model executes a selection of relevant attributes using data mining technique and integrates a CBR system to retrieve similar cases from an archive of cases of patients successfully treated with the proposed treatment plan. The results revealed that the proposed approach outperformed other models and the results of previous studies by achieving an accuracy of 91.66%.

Keywords

Case Based Reasoning, Data Mining, Decision Tree, Medical Decision Making Approach, Predictive Model, Road Accident, Selection of Relevant Attributes, Traumatic Brain Injuries.

DOI: 10.9781/ijimai.2024.02.001

I. Introduction

Road accidents can have catastrophic physical and psychological effect on individuals, their families, and the community. According to the World Health Organization, road accidents cause 1.35 million fatalities and 20 to 50 million illnesses every year. Eighty percent of road traffic accident victims are children with traumatic brain injury, severe brain damage, and associated chest or abdominal disorders https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. Providing prompt and appropriate first aid, medical diagnosis and treatment to road traffic accident victims in the critical first few hours after the incident significantly increases their chances of survival and reduces the severity of injuries. Doctors may face critical problems when making quick decisions about diagnosing, especially when a child's life is at stake. This study aims to accelerate and improve the management of an injured child by predicting the course of actions for admission to the services of Pediatric Traumatic Brain Injury (PTBI) following the ABCDE protocol that consists of the following: Airway (Voice, Breath sounds); Breathing (Respiratory rate, Chest wall movements, Chest percussion...); Circulation (Skin color, sweating, Capillary refill time...), Disability (Alert, Voice responsive, Pain responsive), and Exposure (Expose skin) [1].

To achieve our objective, we propose a new medical decision-making approach guided by CBR and data mining techniques. The development of the database was entirely carried out by the pilot team of the University Training Research Project entitled "Implementation of a Medical Decision Support System adapted for the care of child victims of accidents". The collection of patients' records were done at three medical institutions in Oran, Algeria: the pediatric surgery clinic and the intensive care unit of the University Hospital Center; and the intensive care unit of Canastel Pediatric Hospital.

The proposed approach includes a multi-step procedure: Collection, preprocessing, selection of relevant attributes by data mining technique, and Case-Based Reasoning (CBR). The effectiveness of CBR relies heavily on the quality of its case-base content [2]. Before incorporating the data collected into Clinical Decision Support Systems' (CDSS) knowledge base [3], it becomes imperative to assess and enhance the data's quality. Consequently, data preprocessing procedures assume primary importance as they lay to improve the accuracy of CBR

Please cite this article in press as:
F. Saadi, B. Atmani, F. Henni, H.Benfriha, Z.Addou, R. Guerbouz. An Effective Prediction Approach for the Management of Children Victims of Road Accidents, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 104-114, 2025, http://dx.doi.org/10.9781/ijimai.2024.02.001

systems [4]. Preprocessing the dataset before prediction is the main concept in this work. This step itself consists of four steps called cleaning, discretization, statistical analysis and classification.

Case-based reasoning (CBR) stands as an essential component of artificial intelligence, involving the resolution of novel problems by drawing upon existing solutions from similar cases [5]. This methodology aligns seamlessly with the problem-solving approach adopted by healthcare professionals when confronted with new cases, making its integration into clinical settings a natural fit. The allure of CBR in medical domains lies in the existence of a comprehensive case base, housing a wealth of patient-specific information, including symptoms, diagnoses, treatments, and outcomes. Consequently, numerous CBR systems have been developed to aid in medical diagnosis and decision support [2].

Designing an effective CBR system hinges on two crucial steps: selection of relevant attributes and case retrieval. The choice of appropriate features for case representation and the effective retrieval method are pivotal in ensuring the quality of CBR [6]. As a result, much of the research on CBR primarily revolves around addressing these specific issues [7].To improve our CBR system, DM techniques are used. The concept "data mining" has become more often employed in the literature on medicine during the past several years [8]. Its use in the processing of medical data, however, has only lately become somewhat more widespread. This is especially true for real-world applications in clinical medicine, which may profit from particular data mining techniques that can perform predictive models, take advantage of the clinical domain's knowledge, and describe proposed decisions after the models have been used to support clinical decisions [8]. The decision tree is integrated in the approach as a discriminating tool that selects the most relevant attributes, removes less significant attributes, and thereby reduces the number of attributes utilized in the similarity calculation, hence speeding up the retrieval phase [9].

The paper is organized as follows: Section II summarizes the relevant works available in the literature. The proposed approach is described in detail in section III. Section IV includes the experimental results obtained and a comparison with other selected research works. Finally, conclusions are provided in Section V.

II. Background

The volume, complexity and dynamics of clinical information are a challenge for doctors. Since the 1960s, researchers have envisioned the day when computers could help build a system that processes medical data like a doctor, offers insight into the nature of medical issue resolution, and permits the building of formal clinical reasoning models, and, most importantly, decreases medical error and misdiagnosis. Any technology that can improve the ability to correctly diagnose human disease is a necessary advancement for the well-being of humanity.

Information technologies have been used in the medical field since computers were invented, and various types of computer applications have been created, including Clinical Decision Making Approaches (CDMA) which have been increasingly popular in several medical fields [10]. According to these authors [10], the CDMA improves healthcare quality by providing more precise diagnoses and treatments that are more effective and reliable, while also lowering healthcare costs and avoiding errors caused by doctors' lack knowledge. The reasoning is a crucial function performed by the CDMA inference engine, which combines medical knowledge with patient-specific data and delivers relevant decisions. Several methods for representing medical reasoning have been developed in Artificial Intelligence (AI) and are employed in the construction of decision support systems. These include rule based reasoning [10], [11], Artificial Neural Network [12], KNN [13], ontology [14], association rules [15] case-based reasoning [16]–[18], decision trees [19], random forest [20], fuzzy logic [21], [22], ...etc. Table I lists some Clinical Decision Support Systems (CDSS), their reasoning techniques, and their application area. CBR has demonstrated to be highly promising in using intelligent systems in the healthcare business, among the various AI reasoning techniques. It is a significant methodology and an area of machine learning [5] used in the creation and enhancement of CDSS in a number of projects involving diverse medical applications [9].The first CDSS used pure CBR, i.e. they used CBR alone as a reasoning methodology, and these systems were successful. But their use in practice remains limited due to the complexity of the medical field which cannot be treated with pure CBR.

Many successful hybridization techniques of CBR with other computing methods have been developed in Artificial Intelligence to try to solve the limitations of classical CBR and to achieve the mission of a CDSS which is the improvement of time and accuracy of medical diagnosis [19].

Retrieve-Reuse-Revise-Retain are the four main steps that make up the CBR cycle, the first step (Retrieve) is the most important and expensive phase in the cycle. In this phase the CBR system searches between past cases the most similar case, or more precisely the case that has the problem part similar to the target case. This research is based on the similarity calculus among cases [9]. To develop an efficient CBR-based system, the complexity of the retrieval step must be optimized and improved. Several research works have experimented the integration of techniques issued from DM to improve the efficiency of this step. Guo and Wu [23] developed a Bayesian Network model and CBR hybrid system for use in the healthcare industry. The goal of this system is to increase case retrieval accuracy in the context of big data. Mansoul & Atmani [24] proposed a new approach destined for the medical field that combines CBR and multi-criteria analysis (MCA). This combination aims to facilitate the choice of a solution among several solutions proposed in the retrieval step; which improves this phase. Benfriha et al. [25] suggested an approach for the initial health care of a child's PTBI in the traffic accident. The approach's goal is enhance the retrieval phase in the CBR cycle by using multi-label text categorization which allows reducing the search space and keeping only the most relevant cases and therefore speeding up the retrieval step. The retrieval is related to the representation of the data and to the similarity measures used. Indeed, two aspects of retrieval can quickly become time-consuming, particularly when it comes to solving real problems: the case's basic size and the large number of features. Many works on CBR has proposed improving the retrieval phase through the reorganization of the case base. Mansoul & Atmani [26] proposed a CBR system for predicting the best presumptive diagnosis of orthopedic illnesses. During the retrieval process, the suggested systems include clustering to decrease the base case. Malviya et al. [27] created a medical image retrieval system named (CBMIR) for finding and recovering lung computed tomography (CT) images from a large medical image dataset; this system combines CBR with k-means clustering-based segmentation. Saadi et al. [28] integrated fuzzy clustering in the retrieval phase of CBR. The goal of this work was to minimize the case base and improve the retrieval step by optimizing the similarity computation time. Benamina et al. [29] combined a fuzzy decision tree and CBR to realize a Fuzzy CBR medical system for diabetic patients. The objective of this combination is to improve the retrieval phase by reducing the complexity of the similarity calculation.

The amount of attributes clearly influences the similarity calculus. The more attributes, the higher the processing time is. To reduce the number of attributes, several DM techniques were proposed for

TABLE I. Medical Systems Integrated Data Mining Techniques in CBR Cycle

Reference	System	Technique(s) used	Application field
[10]	-	Rule-Based Reasoning CBR	Management of Drugs Intoxications in Childhood
[11]	-	Rule-Based Reasoning	Asthma diagnosis
[12]	-	ANN	Classification of causes for non- adherence with medication
[13]	-	K-Nearest Neighbor (KNN	Prediction the diabetes disease
[14]	-	Ontology	Depression
[15]	-	Bayesian Network association rules	Diagnosis Vaccination: Detection of lost ones and abundance causes in relation to the mother's socio-economic characteristics.
[16]	Protos	CBR	Classification and Diagnosis Hearing disorders
[17]	Electronic Medical Record system (ArdoCare)	CBR	Electronic healthcare suspicion of aortic pathology
[18]	CBR-Nursing Care Plans System (CNCPS)	CBR	Formulation of effective nursing care plans
[19]	-	Decision trees	Classification of diabetes, hepatitis and heart diseases
[20]	-	Random Forests Classifier	Diagnosis of Heart Arrhythmia
[21]	CKD (Chronic Kidney Disease) diagnostic fuzzy expert system	Fuzzy logic	Prediction chronic kidney disease
[22]	-	Fuzzy C-Means (FCM) clustering	Prediction of Chronic Kidney Disease Progression

selecting relevant variables for the similarity calculation in order to speed up the retrieval step. As a result, the relevant variables are chosen with the goal of reducing the number of attributes used to measure similarity, allowing decreasing the computation time and accelerating and improving the retrieval phase and model performance, such as predictive accuracy [30]. Feuillâtre et al. [31] developed a clinical decision tree for the selection of attributes in order to optimize the retrieval step in CBR. Jarmulak et al. [32] applied the genetic algorithms for the selection of relevant features. Ayed et al. [33] eliminated noisy and redundant attributes and left only the relevant ones for the application of similarity measures by using the Relational Evidential C-Means (RECM). Addisu et al. [34] designed a CBR diagnostic framework for malaria diagnosis and applied an information gain algorithm to select relevant attributes. Henni et al. [35] enhance the retrieval step by reducing the number of attributes using the random forest algorithm. Saadi et al. [9] proposed an approach guided by CBR and decision tree. This work integrates a decision tree to eliminate irrelevant attributes used in the similarity calculation.

III. The Proposed Predictive Approach (TBI-CDMA)

In this article, we propose a new Clinical Decision Making Approach (CDMA) based on CBR, intended for the management of children victims of traffic accidents and who have suffered a traumatic brain injury. This approach is adopted for developing an efficient predictive model (TBI-CDMA) to predict the admission procedure of pediatric traffic accidents victims in pediatric intensive care units. We apply this approach to a real dataset that we have collected following the ABCDE protocol. ABCDE is a protocol for assessing and treating critically ill or wounded people in a systematic way [1]. The Protocol can be used in any clinical emergency, in the street without any equipment or, in a more sophisticated form, in emergency rooms, general wards of hospitals and critical care units. It was developed by the American College of Surgeons in the USA to improve the management of polytrauma victims by early detecting physiological changes that put them in danger of mortality [36]. Since the PTBI dataset collected represents only the signs and symptoms existing in ABCD, the experts preferred to remove the gestures associated with Exposure (E) and followed the ABCD protocol. Fig. 1 summarizes the essential phases of the proposed system.

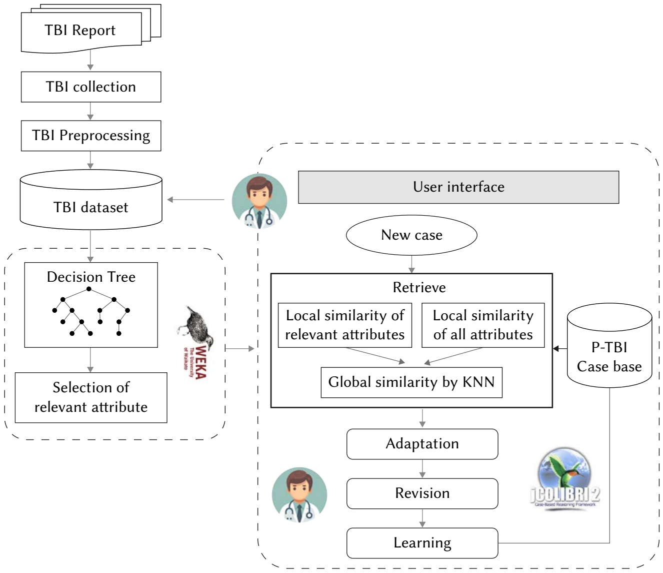


Fig. 1. Overview of TBI-CDMA.

This section presents the steps described in Fig. 1.

A. PTBI Collection

The term "data collection" refers to a systematic approach that entails gathering and measuring many types of information in order to provide a comprehensive and precise picture of an area of interest. The collection of data allows us to answer pertinent questions, evaluate results, and better predict future probabilities and trends. In the intensive care unit at the study site, most of the clinical and administrative information circulating in all the health services is still kept on paper. For this study, we used clinical reports from child traffic accident victims. Since this is the first study in this area, we have selected the records of patients (children) who present a head injury.

B. PTBI Preprocessing

Prior to constructing a model, the data requires preprocessing to derive usable variables from raw data. This crucial step involves extracting meaningful variables and values from the data, and the expertise of healthcare professionals plays a significant role in this process. While removing missing values and outliers generally enhances algorithm accuracy, it is essential to recognize that missing data can carry valuable information, a discernment only achievable by a healthcare professional. For this fact, the main idea in our work is preprocessing the dataset before making the prediction. The preprocessing stage is divided into four sections: cleaning, discretization, statistical analysis and classification. The management of noise or missing values will be performed on the collected dataset as a first preprocessing step. Then, a step of discretization of continuous values into a categorical representation to preserve the learning efficiency and the relevance of the model to build [37], knowing that this step was done with the help of experts who gave us the cut-off points for the attributes. The next step consists in a statistical analysis on the variables by applying a descriptive analysis for each variable and evaluating the correlation and relation between features (attributes) using the Mann-Whitney U test (or Wilcoxon rank-sum test) [38], and finally an evaluation of some classifiers on this dataset is mandatory to validate PTBI. The confusion matrix in Table II is used to compute the performance metrics that are employed in the evaluation.

- Accuracy: is the rate of correct classification, it's defined by (1):

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{1}$$

- Precision: the frequency with which the classifier correctly classifies all of the predictions, as shown in (2) (1):

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

- Recall: Proportion of actual positive results correctly identified, as seen in (3):

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

- F-Measure: is calculated based on precision and recall, as shown in (4):

$$F - Measure = \frac{2 * TP}{2TP + FP + FN} \tag{4}$$

- Kappa Statistic: is a measure of inter-rater agreement for categorical or nominal data. It quantifies the level of agreement between two or more annotators (observers) in the context of classification or labeling tasks. Kappa values range between -1 and 1, where 1 indicates perfect agreement, 0 indicates agreement equivalent to random chance, and values less than 0 signify less agreement than expected by chance..
- ROC: is the ratio between the recall (True Positive Rate TPR (3)) and the FPR (False Positive Rate defined in (5)):

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

- Specificity: is the rate of true negatives defined by (6):

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

TABLE II. Format of Confusion Table

	Predicted Class			
Actual class	True	Positives	False	Positives
	(TP)		(FP)	
	True	Negatives	False	Negatives
	(TN)		(FN)	

C. The Selection of Relevant Attributes

The success of CBR (Case-Based Reasoning) relies on the quality of data (cases) and the speed of the retrieval process, which can be time-consuming, especially when dealing with a large number of cases. Consequently, a case may contain a significant number of attributes [39]. Some attributes are crucial for case reasoning, while others may be unnecessary and result in increased complexity during the retrieval phase. To ensure a robust CBR system performance, it becomes essential to manage the content of the case base. One of the proposed solutions to address these issues is attribute selection. There are several methods of selection of relevant attributes like Chi-square, Euclidean distance, T-test, information gain, correlation-based feature selection, decision tree, ...etc.

Some classification algorithms have earned a reputation for being able to focus on relevant attributes while ignoring others that aren't. The decision tree is a widely used technique for attribute selection in machine learning and data mining [40]. This technique involves building a simple structure where non-terminal nodes represent tests on attributes, and terminal nodes reflect decision outcomes. The test attribute at each node is determined using the information gain measure, which tends to favor attributes with a larger number of values. This algorithm which is an extension of the ID3 algorithm (the basic decision tree induction algorithm) was further improved by C4.5 [41], calculates the rate of information gain (Gain Ratio) at the place of information gain. The C4.5 tree utilizes the gain ratio to determine splits and select the most important attributes. Consequently, the C4.5 decision tree excels at focusing on relevant features while disregarding irrelevant ones. Compared to other strategies, C4.5 generates fewer rules, resulting in reduced error rates and higher precision in the result set. Additionally, C4.5 constructs a trimmed tree, ensuring faster results compared to alternative techniques. The Gain Ratio decreases and corrects the information gain bias by taking the intrinsic information of a split into account. It is therefore weighted by a function that penalizes tests that split the elements into too many sub-classes. This distribution measure is named SplitInfo. The equations (7, 8) refers to the Gain Ratio [42]:

$$GainRatio = \frac{InformationGain}{SplitInfo} \tag{7}$$

or

$$GainRatio = \frac{Entropy(before) - \sum_{j=1}^{k} E\,ntropy(j, after)}{\sum_{j=1}^{k} w_j\, log_2 w_j} \tag{8}$$

Where *before* is the dataset before splitting the data. *K* is the number of subsets generated by the splitting, (*j*, *after*) subset *j* after splitting the data.

$$Entropy(S) = \sum_{i=1}^{c} -p_i log_2 p_i \tag{9}$$

S is the current state, p_i is the probability of an event *i* of state *S*.

The attribute with the highest gain ratio is chosen as the splitting attribute [43]. In the decision tree generated, non-leaf nodes are deemed relevant attributes. By employing the C4.5 algorithm, the attributes resulting from the decision tree become the most significant ones, and these will be used in the similarity calculation, instead of using all attributes. This approach ensures that only the most informative attributes are considered for accurate similarity calculation.

D. Case Based Reasoning

The central notion in a CBR system is the case that represents in our dataset a patient. A record in the PTBI dataset is composed of descriptors divided into a problem part and a solution part. At this stage, we have a case base ready to use in the CBR system. The cycle begins with the retrieval step. In order to succeed in this phase, which is based on the similarity calculation, we need to identify the best similarity measures that allow us to produce good results. Similarity refers to a measure that indicates the strength of a relationship between two features or objects. The purpose of conducting a similarity analysis is to compare two lists of components and calculate a single numerical value representing their evaluation [44]. The similarity measurements are instrumental in retrieving similar cases from the case-based approach.

Among the various distance measurement methods, Euclidean distance is the most commonly used, relying on the spatial location of objects. This method computes a distance as the square root of the sum of squares of numerical differences between two analogous objects. The standard Euclidean distance serves as the fundamental approach for describing the relationship between two cases, forming the neighboring figures of an arbitrary case [45]. In CBR, The Euclidean distance stands as the most prevalent distance metric used to measure the similarity of numerical data [46], and the equality distance which is a simple function for categorical data that have fixed values [47]. Thus, we have exploited these local similarity functions (between the attributes of the cases) and implemented them in the jCOLIBRI platform. The equations for these functions are:

Euclidean Distance: the formula of the euclidean distance is defined by (10):

$$D(x,y) = \frac{\sqrt{x^2 - y^2}}{x + y} \tag{10}$$

Equal: if two attributes (x,y) are equal it returns 0, otherwise it returns 1.

After computing the similarity measures between attributes, the global similarity calculates the distance between two cases using the K-Nearest Neighbor (KNN) technique, which is widely employed in Case-Based Reasoning systems. The algorithm computes the similarity between a new case and a stored case by performing a weighted summation of the pairwise attribute similarities.

With the normalization of local similarity functions, the resulting global similarity value lies within the range of 0 to 1, where 0 indicates complete similarity between the cases.

IV. Experimentation and Results

The PTBI dataset collected from the services mentioned before will be used for the prediction of the course of action to be taken on admission to the services concerned. Most collected clinical information was stored in paper format and recorded in chronological order in an archived register. PTBI was collected in the period between 2017 and 2021. This was done using a Java application that captures the information of patients who have traumatic brain injury. Four parts make up the preprocessing step:

- Cleaning: The data cleaning process was essential to ensure the dataset's quality and reliability. Various issues, such as empty, duplicated, unreadable files, and missing values, were identified during a thorough examination. Empty files were promptly addressed by consulting with doctors and deleting them to avoid any misleading results. Duplicated files were carefully managed, keeping only one representative instance to maintain efficiency and consistency in the analysis.
- Discretization: During the discretization process, several variables were transformed into discrete categories or intervals. This was done for variables such as systolic pressure, diastolic pressure, heart rate, Glasgow score, and more. The decision to discretize these variables was based on the recommendations provided by doctors who were actively involved in the research project. They provided valuable insights and expertise in determining the appropriate cut points or thresholds for discretization. In addition to the input from doctors, international protocols used in Algeria were also consulted to guide the discretization process.
- Statistical analysis: Before deciding if the dataset is ready for use, we applied a statistical analysis under SPSS software. A descriptive analysis is performed on every attribute and the correlation between attributes was calculated. This statistical analysis allowed us to validate the PTBI and determine the attributes that we will be using in TBI-CDMA. Our dataset must be in the form of a case base, the case represents in our dataset a patient. PTBI comports 174 cases. A case of PTBI consists of 36 descriptors divided into a problem and a solution presented according to the ABCD Protocol. The problem part is composed of five categories. Fig. 2 shows the description of a case.

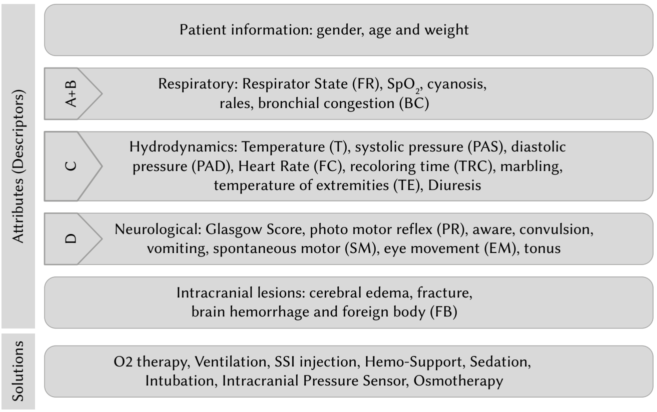

Attributes (Descriptors)
Solutions
Patient information: gender, age and weight
A+B
Respiratory: Respirator State (FR), SpO2, cyanosis, rales, bronchial congestion (BC)
C
Hydrodynamics: Temperature (T), systolic pressure (PAS), diastolic pressure (PAD), Heart Rate (FC), recoloring time (TRC), marbling, temperature of extremities (TE), Diuresis
D
Neurological: Glasgow Score, photo motor reflex (PR), aware, convulsion, vomiting, spontaneous motor (SM), eye movement (EM), tonus
Intracranial lesions: cerebral edema, fracture, brain hemorrhage and foreign body (FB)
O2 therapy, Ventilation, SSI injection, Hemo-Support, Sedation, Intubation, Intracranial Pressure Sensor, Osmotherapy

Fig. 2. Description of Case Base.

The solution part of a case encompasses the course of actions to be executed by the doctor, which is presented in the form of plans. To organize and structure these plans effectively, the ABCD protocol is employed.

However, certain data mining techniques, such as the decision tree, require that the class variable be represented with a single column. In collaboration with health experts, a coding system was devised to codify the different plans into six distinct classes. These classes are described in detail in Table III, with each class listed in descending order based on the severity of the patient's condition upon admission to the pediatric intensive care unit. The organization of these classes allows for effective utilization of DM techniques, such as decision trees, to aid in clinical decision-making processes.

- C5: it is the conduct to hold for a seriously affected case, it includes oxygenation with a nasal probe, non-invasive ventilation, application of an osmotherapy and hemo-support, it must be injected with an isotonic saline, intubated and sedated, and monitored their intracranial pressure.
- C4: designates the actions to be done for patients a little less serious compared to the previous class, these actions are represented by oxygenation with a nasal probe, invasive ventilation, isotonic saline injection, intubation and sedation.
- C3: includes actions to be taken for a patient who does not have a serious neurological problem, so it is enough to give him oxygenation with a mask, non-invasive ventilation, and isotonic saline injection.

TABLE III. Coding of Classes

A+B		C		D				Class
O2 therapy	Ventilation	SSI injection	Hemo-Support	Sedation	Intubation	Intracranial Pressure Sensor	Osmotherapy	
nasal probe	invasive ventilation	✓	✓	✓	✓	✓	✓	C5
nasal probe	invasive ventilation	✓		✓	✓			C4
Mask	non-invasive ventilation	✓						C3
Mask	non-invasive ventilation							C2
		✓						C1
								C0

TABLE IV. Results of the Application of Wilcoxon Rank-sum Correlation Test Between Each Feature and the Target Feature

Attribute	Age	Gender	Weight	PAS	PAD	FC	TE	Diuresis	Marbling	TRC	SPO_2	FR	Cyanosis	RALES	BC	Glasgow	Tonus	Aware	Convulsion	Vomiting	SM	EM	PR	Fracture	CE	FB	BH
P_value	0,31	0,05	0,19	0	0,79	0,45	0,42	0	0	0	0,03	0,91	0	0,62	0	0	0,04	0	0	0,1	0	0	0	0	0	0	0

- C2: designates the course of actions for patients who ot have a neurological disability or a hemodynamic problem but solely have respiration problems, oxygenation using a mask and non-invasive ventilation are the adequate actions.
- C1: represents the solution for a patient who has a small circulation problem, an injection of SSI is enough for him.
- C0: denotes that the patient is in good health and does not require treatment.

The authors utilized a widely used univariate statistical test called the Wilcoxon rank-sum test. This test generates a p-value, which represents the probability of a feature being associated with the target variable. The resulting p-value serves as a score that facilitates the creation of a ranking, highlighting the attributes (features) based on their level of correlation with the target variable.

A low p-value, close to 0, indicates a strong relationship between the examined attribute and the class variable. Consequently, such attributes are considered highly relevant in the context of the analysis. On the other hand, a high p-value, close to 1, suggests that the attribute is not correlated with the class variable, making it less relevant for the study (Table IV).

By employing the Wilcoxon rank-sum test and analyzing the resulting p-values, the authors were able to identify and rank the attributes based on their degree of linkage (correlation) with the target variable. This approach helped to identify significant attributes that are closely related to the class variable and those that have less relevance in the analysis [48]. Upon referring to Table IV, it becomes evident that the variables highlighted in grey are strongly correlated (p-value=0) with the class variable, which represents the action plan implemented by the doctor. The shading of the grey color varies, with a darker shade indicating a higher correlation coefficient, while a lighter shade suggests a decrease in the correlation strength. In contrast, cells without any color signify variables that are not correlated with the class variable.

- Classification: The dataset is evaluated using Bagging (REPTree), NB, PART, J48, RF, SVM, and KNN classification techniques. Table V lists the results of the different evaluation measures.

  According to the data presented in Table V, the evaluation measures in terms of Accuracy, F-Measure, Precision, Recall, Kappa, and ROC Area exceeded 89%. These impressive results provide encouraging evidence, validate the reliability of our dataset, and provide strong support for the success of our research study.

TABLE V. Evaluation Measure of Dataset

Classifier	Accuracy	Precision	Recall	FMeasure	Kappa	ROC Area
Bagging	92.52	0.92	0.92	0.91	0.89	0.99
BN	93.10	0.93	0.93	0.93	0.89	0.99
PART	94.82	0.94	0.94	0.95	0.92	0.99
SVM	98.27	0.98	0.98	0.98	0.97	0.99
J48	94.25	0.93	0.94	0.93	0.91	0.99
KNN (IBK)	100.00	1.00	1.00	1.00	1.00	1.00
RF	100.00	1.00	1.00	1.00	1.00	1.00

PTBI case base was randomly split into train and test sets (70%, 30%). In order to obtain better results, several decision trees are generated under the Weka, Orange, Tanagra, Sipina and R platforms. The decision trees obtained are similar. They have four levels, except the Weka platform which generated a tree of 5 levels. We opted for the decision tree (Fig. 3) implemented by the algorithm C4.5 (J48) under the WEKA platform because the attributes it generated are the most logical. C4.5 select the attribute having the maximum informational gain rate value for splitting the node.

The results obtained from the decision tree analysis (Fig. 3) demonstrated the significant benefits of utilizing a DM technique. By employing the decision tree, a substantial reduction in attributes was achieved. Specifically, the decision tree identified and listed seven highly relevant attributes with the highest informational gain rate.

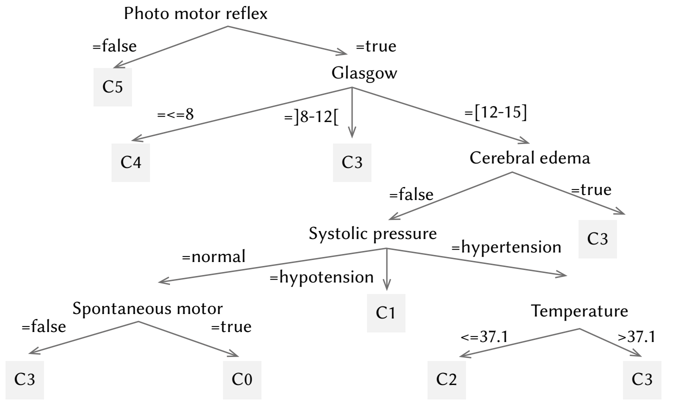


Fig. 3. Decision tree generated by WEKA.

As mentioned previously, conducting a statistical study on PTBI was crucial to assess the correlation between various attributes and the target class. Interestingly, the attributes generated by the decision tree were found to be among those exhibiting strong correlations with the class variable. These attributes proved to be instrumental in predicting the class, thereby confirming the effectiveness of our model. The validation of these results was further supported and endorsed by our collaborating doctors. Their expertise and insights reinforced the credibility and reliability of the model's outcomes, validating the relevance and importance of the selected attributes for predicting PTBI. Overall, the implementation of the decision tree and the statistical study offered valuable insights into the significance of certain attributes in the context of PTBI prediction.

After the generation of the case base and the selection of the most relevant attributes, it is now possible to model the TBI-CDMA solution in the framework jCOLIBRI.

- The case representation: In the context of CBR systems, there is no standardized representation for a case, leading us to adopt a vector form representation to effectively capture and organize the information collected and available in the files. The case is structured into two distinct parts: the problem part, which pertains to the PTBI situation. it encompasses crucial details such as the clinical signs, symptoms, and intracranial lesions related to PTBI, providing essential insights into the patient's condition and medical history. On the other hand, the solution part outlines the specific physical gestures applied as a course of action to address the identified PTBI situation. Within these parts, a total of 29 descriptors are included to comprehensively account for the relevant information.
- The retrieval phase is the cornerstone of a CBR system, where the system identifies the most relevant cases by calculating their similarity to the target case stored in the case base. To compute local similarities of the attributes, two similarity measures are employed based on the attributes revealed by the decision tree (refer to Fig. 3). For numerical attributes, the Euclidean distance is used to measure similarity, enabling the system to quantify the similarity between the target case and other cases stored in the case base. For categorical attributes with fixed values, the equal function is applied to assess their similarity. This function facilitates the comparison of categorical attributes by determining whether their values are the same or not.

  The global similarity calculates the distance between two cases in the jCOLIBRI platform. This platform follows the principle of the K-Nearest Neighbor (KNN) method. In this context, the value of k for the k-nearest neighbor is set to 3, meaning that the system considers the three most similar cases to the target case. After identifying the three nearest cases, a majority vote is applied to determine the appropriate gesture that should be taken to save the child's life. This voting process involves considering the actions taken in the three nearest cases and selecting the gesture with the highest frequency as the recommended course of action for the current case.
- Adaptation: After retrieval, the system evaluates the degree of similarity of the selected cases with the current case. The degree of similarity determines whether an adaptation is necessary or if the solution can be used as it is. If the solution requires adaptation, the doctors are involved in adapting the therapy for the child victim of a road accident. Concerning this phase, it's common in the majority of CBR-based systems to entrust this task to domain experts due to the absence of predefined adaptation rules. Furthermore, the field of medical diagnosis, which revolves around the well-being of patients, is highly sensitive. In such domains, manual adaptation is viewed as a positive rather than a negative aspect. Additionally, as emphasized in [49], the authors assert that "when the knowledge domain is ambiguous or lacks clarity, the development of automatic adaptation mechanisms becomes challenging or is discouraged."
- Revision: The doctor validates the adapted solution.
- Learning: If the generated solution does not exist in the case base and is considered successful in the review state, this new case is added to the case base allowing the system to learn.

A. Performance Evaluation

Our TBI-CDMA system is designed to assist doctors in intensive care services by predicting the appropriate course of action for managing children who have been victims of road accidents. To assess the system's performance, we applied it to the test cases, which constituted 30% of the case base. We selected five real examples (target cases) from the test set, as outlined in Table VI. These target cases were then subjected to the TBI-CDMA analysis to facilitate a comparison between the results obtained from the retrieval process using all attributes and the retrieval process using only the relevant attributes derived from the decision tree. By conducting this comparison, we aim to evaluate the impact of attribute selection on the system's performance and determine the effectiveness of the decision tree in identifying the most crucial attributes for making accurate predictions. This analysis will provide valuable insights into the system's ability to deliver precise and relevant recommendations, enhancing its overall utility in real-world medical scenarios.

Table VII illustrate the result of applying the TBI-CDMA with all attributes and only the relevant attributes for all target cases given in Table VI.

As demonstrated in Table VII, the TBI-CDMA system achieved successful predictions of the best plan of action for case 3 when utilizing all attributes. However, when employing only the relevant attributes derived from the decision tree, the system achieved successful predictions of the best course of actions for 3 out of 5 cases. The performance of the TBI-CDMA system in accurately predicting the best course of actions highlights the effectiveness of the decision tree in identifying important attributes, leading to improved decision support and more informed medical management for children victims of road accidents. According to the doctors' assessment, TBI-CDMA exhibits tolerable errors, with most incorrectly classified cases still providing applicable plans (gestures that can be performed). For instance, in case 4, the predicted class (C3) includes additional actions like an injection of isotonic saline, despite the actual class being C2, denoting non-invasive ventilation. Similarly, in case 5, the predicted class (C2) includes additional preventive actions like oxygenation and isotonic saline injection, while the actual class is C1, representing only

TABLE VI. Target Cases

Case	Age	Gender	Weight	T	PAS	PAD	FC	TE	Diuresis	Marbling	TRC	SPO_2	FR	cyanosis	RALES	BC	Glasgow	Tonus	Aware	Convulsion	Vomiting	SM	EM	PR	Fracture	CE	FB	BH	Class
1	48	M	20	36.9	Normal	Normal	Bradycardie	Cold	Preserved	false	3	Normal	Normal	false	Normal	false	[12,15]	Normal	true	false	false	true	Normal	true	false	false	false	false	C3
2	164	M	50	37.5	Normal	Normal	Bradycardie	Cold	Preserved	false	3	Normal	Normal	false	Normal	false	< = 8	Hypo	true	false	false	false	false	true	false	false	false	false	C4
3	156	M	50	38.7	Hypertesion	Normal	Bradycardie	Hot	Preserved	false	ă3	Normal	Normal	false	Normal	false	< = 8	Hypo	False	False	False	false	false	False	True	True	False	True	C5
4	160	M	50	36.4	Hypertesion	Normal	Bradycardie	Normal	Preserved	false	ă3	Normal	DR-H	false	Normal	false	[12- 15]	Normal	True	False	False	True	Normal	True	False	False	False	False	C2
5	64	F	18	37	Normal	HVS-H	Bradycardie	Normal	Preserved	false	ă3	Normal	Normal	false	Normal	false	[12- 15]	Normal	True	False	True	True	Normal	True	False	True	False	False	C1

TABLE VII. Results Obtained by TBI-CDMA

Actual class of cases		Predicted class with relevant attributes	Predicted class with all attributes
Case 1	C3	C3	C4
Case 2	C4	C4	C5
Case 3	C5	C5	C5
Case 4	C2	C3	C0
Case 5	C1	C2	C3

isotonic saline injection. The doctors view these additional actions as preventive measures that can be applied to any patient admitted to the pediatric intensive care unit, regardless of their health status. Furthermore, the system empowers doctors to evaluate the proposed plan's suitability for a new case. If the plan is deemed appropriate, it can be directly applied; otherwise, modifications can be made as needed. This flexibility and control provided to the doctors enhance the system's usability and efficacy in real-world medical scenarios. Overall, the TBI-CDMA system has delivered encouraging results, with plans validated and approved by the doctors. Its ability to provide relevant and applicable plans, along with the option for doctor intervention and modification, underscores the potential of the system to enhance medical decision-making and improve patient care in pediatric intensive care units. Performance measurements can be used to assess the efficacy of our TBI-CDMA, we calculated the well-known performance and evaluation measures: the accuracy (equation 1), the precision (equation 2), the sensitivity (recall) (equation 3), and the specificity (equation 7) which are the common statistics to evaluate a model's performance. A comparison of the proposed approach (Approach 1) with other approaches was done (Table VIII):

- Approach 2: The authors of this approach [25] used the multi-label techniques to enhance the retrieval step. Their experiments were conducted using PTBI case base.
- Approach 3: in this approach, Informational gain is used for the selection of relevant attributes. Nine attributes are selected as relevant attributes for the similarity calculation, and then retrieval step by KNN has done.
- Approach 4: retrieval by KNN using all attributes.

TABLE VIII. Evaluation of the Proposed Approach

Measure	Approach 1	Approach 2	Approach 3	Approach 4
Precision	75.00	71.15	69.23	65.38
Sensitivity	75.00	71.15	69.23	65.38
Accuracy	91.66	90.38	89.74	88.46
Specificity	95.00	94.23	93.84	93.07

As depicted in Table VIII, the proposed approach (Approach 1) exhibits superior performance compared to the other three approaches. Approach 1 achieves an accuracy score of 91.66, precision and sensitivity of 75.00, and specificity of 95.00, outperforming the alternatives. Approaches 3 and 4 also demonstrate good performances, while Approach 1 fares the worst for all measures.

The performance of the proposed model was extensively evaluated using seven medical datasets obtained from the UCI Machine Learning Repository [50]. These datasets are well-regarded for their application in benchmarking and real-world data collection within the CBR and data mining community. Our model was applied to these datasets to validate its effectiveness. Each medical dataset was specified with its attributes and instances. Table IX presents comprehensive details about the datasets, including accuracy and Roc area results, showcasing the performance of our model when applied to these datasets.

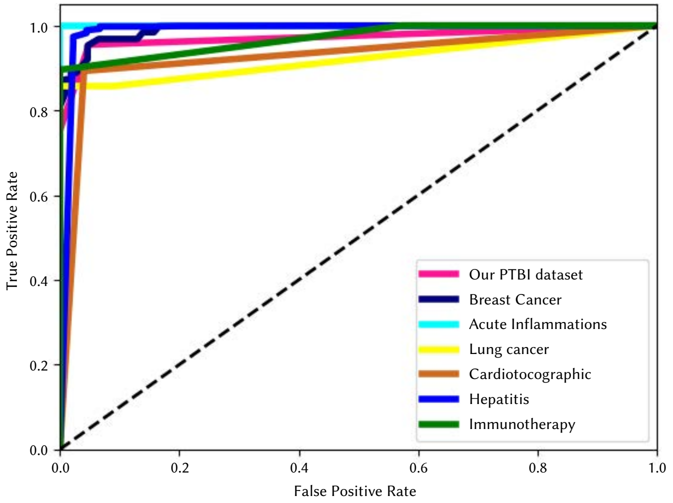


Fig. 4. ROC Curves.

As demonstrated in Table IX, our model that integrates the decision tree in the retrieval step to ameliorate this step and the CBR process has achieved remarkable levels of accuracy, surpassing 90%, when applied to the medical datasets. Furthermore, the ROC area for all datasets is notably high, as depicted in Fig. 4. Our model exhibits the highest classifier performance not only for our specific dataset but also for other publicly available datasets, thus validating its efficacy and reliability.

These findings suggest that employing pure CBR alone may not be sufficient to design a Clinical Decision Support System (CDSS) with optimal performance. Instead, integrating data mining (DM) techniques in the retrieval phase can enhance the overall performance of a CBR-based system. The incorporation of relevant attributes revealed by the decision tree in TBI-CDMA contributes to improved retrieval and surpasses the traditional CBR approach, enhancing its potential as an effective tool for assisting doctors in making critical decisions regarding the management of pediatric patients involved in road accidents.

V. Conclusion

In this work, the authors proposed a novel Clinical Decision Making Approach (CDMA) known as TBI-CDMA, specifically designed for the care of children victims of a road accident, from the site of the accident until their arrival in the pediatric intensive care unit. The authors applied their approach to a real dataset (PTBI) elaborated by them. The proposed TBI-CDMA consisted of a series of multistep procedures, including data collection, preprocessing, selection of relevant attributes, and Case-Based Reasoning (CBR). Real-world datasets often contain incomplete, noisy, and inconsistent data, which can obscure valuable patterns. Hence, the preprocessing step plays a crucial role in generating high-quality data and enhancing the accuracy and efficiency of the models [51]. Within TBI-CDMA, data preprocessing assumed a fundamental role with a significant impact on its accuracy. it involves various processes such as data cleaning, discretization, statistical analysis, and classification. These steps aimed to prove the PTBI dataset, making it more suitable for the subsequent stages of the decision-making process, and ensuring the system's overall effectiveness in providing appropriate and timely care for the child victim of road accident. In the statistical analysis step, an evaluation of the correlation of the features (attributes) with the target was done; and in the classification step, we evaluated some classifiers on our dataset. The subsequent step involved integrating DM techniques with CBR, especially in the retrieval step. Medical databases are known to be voluminous, and each case can contain numerous attributes. While some attributes are crucial, others may be unnecessary and lead to increased complexity during case retrieval. To ensure optimal performance of a CBR system, maintaining the content of the case base is essential. One proposed solution to address these issues is attribute selection. By identifying and retaining only relevant attributes, the selection of relevant attributes aims to reduce the size of the case base and minimize retrieval time which improved the performance of this step. This approach was tested using other datasets and yielded good results.

Based on the obtained results, the preprocessing step successfully validated the reliability of our PTBI dataset and identified attributes strongly correlated with the treatment plan applied by doctors. This finding further confirms the significant reduction of attributes achieved through the decision tree. Also, our approach demonstrates that utilizing only relevant attributes in similarity calculations generates accurate plans endorsed by doctors, resulting in superior performance compared to other techniques. Overall, the innovative TBI-CDMA approach proved highly effective in preprocessing and validating the PTBI dataset, resulting in faster and enhanced retrieval and medical management, vital for time-sensitive decision-making scenarios. This significant advancement holds great promise for improving the care of child victims of road accidents.

A limitation of the CDMA in relation to the PTBI dataset is that the number of cases collected is reduced, resulting in the dataset not representing all of the signs, symptoms, and gestures applied to a child who has been injured. Once we have more cases, we may be able to improve the model's predictive performance. On another level, there is still work to be done at the core of the model. Indeed, this project did not take into consideration the adaptation of the solutions and was satisfied with a manual adaptation made by the doctors. It is legitimate to think about an automatic or semi-automatic adaptation once the solution is deployed. knowing that in this step, the system evaluates the degree of similarity between the selected cases and the current case, providing the doctors with the most relevant cases along with their respective degrees of similarity. The doctor then chooses the case that they find most appropriate or may make modifications to their solution if they deem it necessary. The idea is to record the modifications made by the expert on the proposed solution in order to generate rules that capitalize on this expertise. This can only be done when the solution is used over a sufficiently long period.

References

[1] T. Thim, N. H. V. Krarup, E. L. Grove, C. V. Rohde, B. Løfgren, "Initial assessment and treatment with the airway, breathing, circulation, disability, exposure (abcde) approach," *International journal of general medicine*, vol. 5, p. 117, 2012.

[2] P. Andritsos, I. Jurisica, J. I. Glasgow, "Case-based reasoning for biomedical informatics and medicine," *Springer Handbook of Bio-/Neuroinformatics*, pp. 207–221, 2014.

[3] S. S. R. Abidi, S. Manickam, "Leveraging xml-based electronic medical records to extract experiential clinical knowledge: An automated approach to generate cases for medical case-based reasoning systems," *International Journal of Medical Informatics*, vol. 68, no. 1-3, pp. 187–203, 2002.

[4] K. C. A. D. Borges, I. F. de Barcelos Tronto, R. de Aquino Lopes, J. D. S. da Silva, "A methodology for preprocessing data for application of case based reasoning," in *2012 XXXVIII Conferencia Latinoamericana En Informatica (CLEI)*, 2012, pp. 1–8, IEEE.

[5] N. Choudhury, S. A. Begum, "A survey on case- based reasoning in medicine," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 8, pp. 136–144, 2016.

[6] F. Saadi, B. Atmani, F. Henni, "Improving retrieval performance of case based reasoning systems by fuzzy clustering," *International Journal of*

Interactive Multimedia and Artificial Intelligence, (2023), doi: http://dx.doi.org/10.9781/ijimai.2023.07.002.

[7] Y. Guo, J. Hu, Y. Peng, “Research of new strategies for improving cbr system,” *Artificial Intelligence Review*, vol. 42, pp. 1–20, 2014.

[8] R. Bellazzi, B. Zupan, “Predictive data mining in clinical medicine: current issues and guidelines,” *International journal of medical informatics*, vol. 77, no. 2, pp. 81–97, 2008.

[9] F. Saadi, B. Atmani, F. Henni, “Integration of datamining techniques into the cbr cycle to predict the result of immunotherapy treatment,” in *2019 International Conference on Computer and Information Sciences (ICCIS)*, 2019, pp. 1–5, IEEE.

[10] B. N. Barigou, F. Barigou, C. Benchehida, B. Atmani, G. Belalem, “The design of a cloud-based clinical decision support system prototype: management of drugs intoxications in childhood,” in *Research Anthology on Decision Support Systems and Decision Management in Healthcare, Business, and Engineering*, IGI Global, 2021, pp. 387–409.

[11] L. J. Hoeksema, A. Bazzy-Asaad, E. A. Lomotan, D. E. Edmonds, G. Ramírez-Garnica, R. N. Shiffman, L. I. Horwitz, “Accuracy of a computerized clinical decision-support system for asthma assessment and management,” *Journal of the American Medical Informatics Association*, vol. 18, no. 3, pp. 243–250, 2011.

[12] S. Walczak, S. R. Okuboyejo, “An artificial neural network classification of prescription nonadherence,” *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, vol. 12, no. 1, pp. 1–13, 2017.

[13] E. K. Hashi, M. S. U. Zaman, M. R. Hasan, “An expert clinical decision support system to predict disease using classification techniques,” in *2017 International conference on electrical, computer and communication engineering (ECCE)*, 2017, pp. 396–400, IEEE.

[14] Y.-S. Chang, C.-T. Fan, W.-T. Lo, W.-C. Hung, S.-M. Yuan, “Mobile cloud-based depression diagnosis using an ontology and a bayesian network,” *Future Generation Computer Systems*, vol. 43, pp. 87–98, 2015.

[15] F. Z. Benhacine, B. Atmani, F. Z. Abdelouhab, “Contribution to the association rules visualization for decision support: A combined use between boolean modeling and the colored 2d matrix,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 5, pp. 38–48, 2019.

[16] E. R. Bareiss, B. W. Porter, C. C. Wier, “Protos: An exemplar-based learning apprentice,” in *Machine learning*, Elsevier, 1990, pp. 112–127.

[17] C. Cândea, G. Cândea, Z. B. Constantin, “Ardocare - a collaborative medical decision support system,” *Procedia Computer Science*, vol. 162, pp. 762–769, 2019.

[18] V. Tang, P. K. Y. Siu, K. L. Choy, H. Y. Lam, G. T. S. Ho, C. K. M. Lee, Y. P. Tsang, “An adaptive clinical decision support system for serving the elderly with chronic diseases in healthcare industry,” *Expert Systems*, vol. 36, no. 2, p. e12369, 2019.

[19] D. S. Kumar, G. Sathyadevi, S. Sivanesh, “Decision support system for medical diagnosis using data mining,” *International Journal of Computer Science Issues (IJCSI)*, vol. 8, no. 3, p. 147, 2011.

[20] E. Alickovic, A. Subasi, “Medical decision support system for diagnosis of heart arrhythmia using dwt and random forests classifier,” *Journal of medical systems*, vol. 40, no. 4, pp. 1–12, 2016.

[21] A. Yadollahpour, J. Nourozi, S. A. Mirbagheri, E. Simancas-Acevedo, F. R. Trejo-Macotela, “Designing and implementing an anfis based medical decision support system to predict chronic kidney disease progression,” *Frontiers in physiology*, p. 1753, 2018.

[22] F. Hamedan, A. Orooji, H. Sanadgol, A. Sheikhtaheri, “Clinical decision support system to predict chronic kidney disease: A fuzzy expert system approach,” *International journal of medical informatics*, vol. 138, p. 104134, 2020.

[23] Y. Guo, K. Wu, “Research on case retrieval of bayesian network under big data,” *Data & Knowledge Engineering*, vol. 118, pp. 1–13, 2018.

[24] A. Mansoul, B. Atmani, “Combining multi-criteria analysis with cbr for medical decision support,” *Journal of Information Processing Systems*, vol. 13, no. 6, pp. 1496– 1515, 2017.

[25] H. Benfriha, B. Atmani, F. Barigou, F. Henni, B. Khemliche, S. Fatima, A. Douah, Z. Z. Addou, “Improving cbr retrieval process through multilabel text categorization for health care of childhood traumatic brain injuries in road accident,” in *Proceedings of Sixth International Congress on Information and Communication Technology*, 2022, pp. 721–731, Springer.

[26] A. Mansoul, B. Atmani, “Clustering to enhance case- based reasoning,” in *Modelling and Implementation of Complex Systems*, Springer, 2016, pp. 137–151.

[27] N. Malviya, N. Choudhary, K. Jain, “Content based medical image retrieval and clustering based segmentation to diagnose lung cancer,” *Advances in Computational Sciences and Technology*, vol. 10, no. 6, pp. 1577–1594, 2017.

[28] F. Saadi, B. Atmani, F. Henni, “Integration of fuzzy clustering into the case base reasoning for the prediction of response to immunotherapy treatment,” in *Mediterranean Conference on Pattern Recognition and Artificial Intelligence*, 2019, pp. 192–206, Springer.

[29] M. Benamina, B. Atmani, S. Benbelkacem, “Diabetes diagnosis by case-based reasoning and fuzzy logic,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 3, pp. 72–81, 2018.

[30] M. A. Hall, “Correlation-based feature selection for machine learning,” 1999.

[31] H. Feuillâtre, V. Auffret, M. Castro, H. Le Breton, M. Garreau, P. Haigron, “Study of similarity measures for case-based reasoning in transcatheter aortic valve implantation,” in *2017 Computing in Cardiology (CinC)*, 2017, pp. 1–4, IEEE.

[32] J. Jarmulak, S. Craw, R. Rowe, “Genetic algorithms to optimise cbr retrieval,” in *European Workshop on Advances in Case-Based Reasoning*, 2000, pp. 136–147, Springer.

[33] S. B. Ayed, Z. Elouedi, E. Lefevre, “Managing uncertainty during cbr systems vocabulary maintenance using relational evidential c-means,” in *2020 International Multi-Conference on:“Organization of Knowledge and Advanced Technologies”(OCTA)*, 2020, pp. 1–10, IEEE.

[34] E. G. Addisu, A. S. Boltena, S. Y. Amare, “Case-based reasoning framework for malaria diagnosis,” *I. J. Information Technology and Computer Science*, vol. 6, pp. 31-48, 2020.

[35] F. Henni, B. Atmani, F. Atmani, F. Saadi, “Improving coronary artery disease prediction: Use of random forest, feature importance and case-based reasoning,” *International Journal of Decision Support System Technology (IJDSST)*, vol. 15, no. 1, pp. 1–17, 2023.

[36] M. D. Von Atzingen, D. R. C. Schmidt, E. A. P. M. Nonino, “Elaboration and application of an evaluation instrument in the immediate postoperative period, based on the advanced trauma life support protocol,” *Acta Paulista de Enfermagem*, vol. 21, pp. 616–623, 2008.

[37] U. Stańczyk, B. Zielosko, G. Baron, “Discretisation of conditions in decision rules induced for continuous data,” *PloS one*, vol. 15, no. 4, p. e0231788, 2020.

[38] F. Wilcoxon, “Individual comparisons by ranking methods,” in *Breakthroughs in statistics*, Springer, 1992, pp. 196–202.

[39] A. Smiti, Z. Elouedi, “Using clustering for maintaining case based reasoning systems,” in *2013 5th International Conference on Modeling, Simulation and Applied Optimization (ICMSAO)*, 2013, pp. 1–6, IEEE.

[40] H. Sun, X. Hu, “Attribute selection for decision tree learning with class constraint,” *Chemometrics and Intelligent Laboratory Systems*, vol. 163, pp. 16–23, 2017.

[41] J. R. Quinlan, “Induction of decision trees,” *Machine learning*, vol. 1, pp. 81–106, 1986.

[42] A. G. Karegowda, A. Manjunath, M. Jayaram, “Comparative study of attribute selection using gain ratio and correlation based feature selection,” *International Journal of Information Technology and Knowledge Management*, vol. 2, no. 2, pp. 271–277, 2010.

[43] J. K. Han, M. Kamber, “Data mining: Concepts and techniques,” Morgan Kaufmann, 2001.

[44] I. Ragnemalm, “The euclidean distance transform in arbitrary dimensions,” *Pattern Recognition Letters*, vol. 14, no. 11, pp. 883–888, 1993.

[45] J. Ahn, S.-H. Ji, S. J. Ahn, M. Park, H.-S. Lee, N. Kwon, E.-B. Lee, Y. Kim, “Performance evaluation of normalization-based cbr models for improving construction cost estimation,” *Automation in Construction*, vol. 119, p. 103329, 2020.

[46] M. T. Rezvan, A. Z. Hamadani, A. Shalbafzadeh, “Case-based reasoning for classification in the mixed data sets employing the compound distance methods,” *Engineering Applications of Artificial Intelligence*, vol. 26, no. 9, pp. 2001–2009, 2013.

[47] A. Idri, A. Abran, T. Khoshgoftaar, “Fuzzy analogy: A new approach for software cost estimation,” in *International Workshop on Software Measurement*, 2001, pp. 28–29, Citeseer.

[48] D. Chicco, G. Jurman, “Machine learning can predict survival of patients

with heart failure from serum creatinine and ejection fraction alone," *BMC medical informatics and decision making*, vol. 20, no. 1, pp. 1–16, 2020.

[49] F. Henni, *Composition Dynamique de Services Web par Apprentissage Artificiel*. PhD dissertation, Université Oran 1, 2015.

[50] C. B. D. Newman, C. Merz, "UCI repository of machine learning databases," 1998. [Online]. Available: http://www.ics.uci.edu/„mlearn/MLRepository.html.

[51] S. Zhang, C. Zhang, Q. Yang, "Data preparation for data mining," *Applied artificial intelligence*, vol. 17, no. 5- 6, pp. 375–381, 2003.

Fatima Saadi

Fatima Saadi is currently a PhD candidate at the University of Oran 1 and affiliated researcher in Laboratoire d'Informatique d'Oran, Algeria. Her research interests include data mining, case-based reasoning,information retrieval, medical decision support systems, machine learning.

Baghdad Atmani

Baghdad Atmani is currently a Full Professor in Computer Science. His field of interest is artificial intelligence and machine learning. His research is based on knowledge representation, knowledge-based systems, CBR, data mining, expert systems, decision support systems and fuzzy logic.

Fouad Henni

Fouad Henni is a teaching researcher in computer science at Mostaganem university, Algeria. His research interests are in the areas of semantic Web services, artificial intelligence, case-based reasoning and deep learning, with a particular emphasis on applications in medical diagnosis. He is member of the Computer Science and new Technologies Lab (CSTL).

Hichem Benfriha

Hichem Benfriha is a computer science teacher in the Department of Technical Sciences, University of Mascara Mustapha Stambouli, Algeria. He is currently a Research Member of Laboratory of Computer Science of Oran. He is currently a PhD candidate in the Computer Science Department of Oran 1 University (Algeria). He received his Master of Science degree in 2012 from the same university. His research interests focus on CBR, data Mining, text mining, information extraction, information retrieval, natural language processing, machine learning and Multi-label classification areas.

Zakaria Zoheir Addou

Zakaria Zoheir Addou is a pediatric anesthesiologist and critical care physician since 2002 at pediatric hospital in Oran, Algeria. He received his PhD in 2017 at the university of Ahmed Ben bella 1 Oran in medical science. He is also a member of laboratory of pediatric injury in the University of Oran. His field of research is safety of anesthesia outside operating room in children.

Rabah Guerbouz

Rabah Guerbouz is a neurosurgeon at the University Hospital of Oran, Algeria since 2007. He obtained his doctorate in 2018 at the University Ahmed Ben Bella 1 Oran in medical sciences. He is a teacher at the national paramedical school.

Learning Analytics Icons: Easy Comprehension of Data Treatment

Daniel Amo-Filva[1]*, Marc Alier[2], David Fonseca[1], Francisco José Garcia-Peñalvo[3], María José Casañ[3]

[1] La Salle, Ramon Llull University, Barcelona (Spain)
[2] Polytechnical University of Catalonia, Barcelona (Spain)
[3] University of Salamanca, Salamanca (Spain)

* Corresponding author: daniel.amo@salle.url.edu

Received 4 March 2022 | Accepted 21 April 2023 | Published 8 April 2024

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

The Learning Analytics approach adopted in education implies the gathering and processing of sensitive information and the generation of student profiles, which may have direct or indirect dire consequences for the students. The Educational institutions must manage this data processing according to the General Data Protection Regulation, respecting its principles of fairness when it comes to information gathering and processing. This implies that the students must be well informed and give explicit consent before their information is gathered and processed. The GDPR propose the usage of recognizable standardized icons to facilitate a general understanding and awareness of how personal data is deemed to be processed in each application context, like an online course. This paper presents a project that aims to provide a set of icons to inform about the treatment of educational data in the Learning Analytics processes and a survey about the student's comprehension of the icons, their meaning, and implications for their privacy and confidentiality. The result presented is a set of icons ready to be integrated into educational environments that apply Learning Analytics to increase transparency and facilitate the understanding of data processing.

Keywords

Data Treatment, Education, GDPR, Icons, Learning Analytics, Privacy, User Profiling.

DOI: 10.9781/ijimai.2024.04.001

I. Introduction

Data scientists can analyze educational data from different perspectives. On the one hand, educational data can be processed with the unique objective of extracting and discovering behavioral patterns. This process is called Data Mining [1]. On the other hand, educational data can be treated with the ultimate purpose of improving any aspect of the teaching-learning methodology. This process is called Learning Analytics [2], [3]. Therefore, Learning Analytics is an analytical approach that collects, analyzes, and visualizes student data to improve the educational context. Reasons for improvement are the processes of tutoring, evaluation, or even student follow-up [4]. This paper focuses on the possibilities offered by Learning Analytics in the educational community, focusing on the ability of institutions to be transparent and able to fulfill the various challenges it presents, especially those related to the treatment and privacy of students' [personal] data. It is worth going back to the beginnings of Learning Analytics to understand the need for transparency and trust that institutions must convey to students.

It was George Siemens [5], who in 2010 took the first steps in this approach, giving its current name and creating the first discussion groups in Google Groups to reflect on the state of the art and possibilities of educational data analysis. Over time, Siemens' extensive dissemination task managed to transcend the term to the education and scientific communities. Currently, Learning Analytics is a field considered by the scientific community to be of high interest, where a large volume of scientific contributions from different authors from around the world are published, unraveling both its underlying model [2] and its opportunities [6] and even weaknesses [7].

Regarding the education community, the use of data for decision-making is gaining adoption at all levels: those related to teachers-students, institutional and inter-institutional [8]. Learning Analytics' origins focused on supporting Massive Online Open Courses (MOOC) type courses [9], [10]. Eventually, its applications have been adapted to other educational contexts and needs. The first utility of Learning Analytics related to MOOCs and Virtual Learning Environments (VLE) was teaching support and dropout rate diminishing [11], [12]. These MOOC courses have low teaching staff and high enrollments, reaching hundreds of thousands of participants in some cases (hence they are called massive courses). Thereon, its uses in VLEs evolved from reducing dropouts [13] to meet other needs such as improving teaching methodologies, student well-being, or even shaping learning spaces. Over time, the adoption of Learning Analytics has bounced from virtual platforms to physical environments. Multimodal Learning Analytics [14] is the branch of Learning Analytics dedicated to analyzing student behavior in face-to-face context through [connected] sensors.

Please cite this article in press as:
D. Amo-Filva, M. Alier, D. Fonseca, F. J. Garcia-Peñalvo, M. J. Casañ. Learning Analytics Icons: Easy Comprehension of Data Treatment, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 115-126, 2025, http://dx.doi.org/10.9781/ijimai.2024.04.001

In any case, the [virtual and face-to-face] learning processes mediated by Learning Analytics collect data of all possible student interactions and academic performance, both treated as students' behaviors [15]. At the same time, data collection goes beyond what is strictly educational or academic. This additional data, considered as metadata, are complementary and may originate from heterogeneous sources such as social networks or financial data. All this data and metadata collection generate a sensitive context making data fragile in all senses [16]. Consequently, concerns arise regarding confidentiality, privacy, and security of the students' data and, in it extends, about their digital identity [17].

Despite this contextual sensitivity, the adoption of Learning Analytics in the educational context has increased due to different factors, including:

- The rapid transition from classrooms to hyperconnected classrooms.
- Deep classroom integration of connected learning devices.
- Digitization of teaching-learning materials and processes.
- The incorporation of third-party educational technology in the form of apps in the cloud.
- The rapid evolution of educational technologies based on Big Data, Artificial Intelligence, and Machine Learning as facilitators of teaching-learning processes.
- The use of cloud computing to reduce IT infrastructure costs.

However, one must not be enlightened by the rapid evolution of technologies and their promises. With the pandemic and the confinement caused by COVID-19 [18], [19], interest in Learning Analytics has increased in all educational stages and worldwide [20]. At the beginning of the pandemic, the collection and processing of educational data made it possible, in the first instance, to understand how students interacted with VLEs, and in the second instance, to give them the appropriate and necessary support. All this data collected in pre-pandemic, pandemic, and post-pandemic is stored, analyzed, and even shared between institutions and countries. Such data treatment is regulated by the General Data Protection Regulation (GDPR) [21] and other data protection laws of each EU member state. These laws exist since it is necessary to regulate any data processing to avoid improper use. Hence, the educational context must enforce these laws, such as transmitting certain aspects to students as decision-making information, even before registering for any course.

The adoption of Learning Analytics can negatively impact the confidentiality [22], privacy [23], and security [24] of student data, as well as their digital identity [25]. In the worst case, students do not realize it until it is too late, their data being misused [26], shared with third parties [27], leaked [28], or used by algorithms [29] with dire results to the students themselves. Different authors have pointed to this type of problem [17], [30]–[32]. Local technologies have even been proposed, substituting or complementary to cloud computing, to ensure that this environment of mistrust reverts to one of absolute certainty of a secure data environment [16].

In any case, whether using local or cloud technologies, the use of Learning Analytics implies a great responsibility regarding the collection, storage, treatment, and sharing of educational data, especially when the data is from minors. For all the above reasons, strict law enforcement is necessary as a tool for preserving data privacy.

A. General Data Protection Regulation

In 2016, the GDPR was approved, however, its entry into force was not scheduled until two years later, specifically, on May 28, 2018. The GDPR establishes the obligations that contract the entities that process and manage personal data, those that by themselves can identify a person. At the same time, it defines five fundamental rights of citizens before such entities: the right to know, the right to request the data controller, the right to rectify data, the right to delete data, and the right to oppose data processing [33]. These rights allow any person to suspend the data processing, facilitate data portability to third parties, revoke the consent given, or even oppose automatic processing.

The study tackles the right to be informed, which includes other rights such as knowing: the purposes of data use, the period of data conservation, and even if there are automated decisions or profiling. Recital 60 [34] of the GDPR establishes that the interested party must be informed:

- "...of the existence of the processing operation and its purposes".
- "...with any further information necessary to ensure fair and transparent processing".
- "...of the existence of profiling and the consequences of such profiling".
- "...whether he or she is obliged to provide the personal data and of the consequences, where he or she does not provide such data".

Considering the above and the educational context, fast, transparent, and easy to understand forms of information are required to:

- Raise awareness of current student data treatment processes in any educational context.
- Let the students know how their data is treated and for what purpose, preferably previous registration to any course.
- Establish a standardized information system and transmission medium in any educational context for obvious reasons.

B. Icons & Learning Analytics

Recital 60 of the GPDR informs about the possibility of using standardized icons to combine with textual information. The purpose of the complementary use of standardized icons is to give a clear, intelligible, and legible view of the intended processing. Besides, in point 7 of Article 12 of the GDPR, the possibility of an iconified representation of Articles 13 and 14 is exposed. These two articles consider all the information shown at Recital 60. Both Recital 60 and Article 12.7, require that electronically presented icons must be machine-readable, in other words, data structured in formats such as P3P, JSON, or XML should accompany the icons.

The use of the icons can fulfill the three points of the previous section. Icons can inform how the data is processed in Learning Analytics and, therefore, generate awareness to students by facilitating access to this type of information. Consequently, we propose as objectives of the study:

- Design descriptive icons of those parts of the Learning Analytics processes that must be reported to students, and that other authors have not designed in their work.
- Develop and make available a tool such as Creative Commons where any VLE administrator can create the appropriate icon packs to inform students of the different treatments of their data from the LMS itself.

The structure of this document is organized into four sections. All four sections show how our work makes available a set of icons regarding the data treatment in education, where methods such as Learning Analytics [or other kinds of as Academic Analytics or Educational Data Mining] are applied to facilitate its comprehension by students. Section I is the introduction. Section II gives the used methodology and fundamentals to design the icons. Section III exposes the results of the different design phases concerning questionnaires results. Section IV presents the conclusions.

II. Methods

The methodology of the study is mixed and of two phases whose purposes are adjusted to the objectives of the study.

A. Phase 1

We propose to start with a documentary methodology using a qualitative-quantitative approach. The purpose of this phase is to review and understand the work done by other authors regarding the iconifying of Articles 13 and 14, both in general and specifically for Learning Analytics. For this purpose, we designed a basic systematic literature review (SLR) [35], [36] using Web of Science and Scopus as database indexes. However, we begin with a mapping of the context before delving into the review itself. These are the mapping questions:

- MQ1. How many domain-related studies have been published in the last six years?
- MQ2. In which media have the articles been published?
- MQ3. Are there authors in common among the selected articles?

We aim to answer the following research question:

- RQ1. What kind of icons has been created related to GDPR?
- RQ2. What icons have been created related to analytical actions?

The answer to these questions will help understand how many Learning Analytics actions have been iconified and if there are any that remain to be defined.

We establish the following inclusion criteria for the search of manuscripts:

- (IC1) The results must be scientific publications.
- (IC2) The results must contain icons related to legal aspects defined in the GDPR.
- (IC3) The results can refer to icons related to Learning Analytics processes.
- (IC4) Results must be published before the enactment of the GDPR.
- (IC5) The language of the results must be in English or Spanish.
- (IC6) The results must be published in scientific conferences or journals without the need for impact.
- (IC7) The results must have been published through a peer-review process (double-blind).

We establish the following exclusion criteria for the search of manuscripts:

- (EC1) The results are not scientific publications.
- (EC2) The results do not contain icons related to legal aspects defined in the GDPR.
- (EC2) The language of the results is different to English and Spanish.
- (EC3) The results are not published in scientific conferences or journals (with or without impact).
- (EC4) The results are not published through a peer-review process (double-blind).

We define database-related search strings for each database index as:

- Web of Science: “GDPR icon*”
- Scopus: TITLE-ABS-KEY (gdpr AND icons)

We conduct the SLR regarding the PRISMA declaration . Thus, the flow of information through the different phases of the systematic review is shown in the flow diagram available in Fig. 1.

The search returns only 12 results. Only 6 are considered valid after removing duplicates and non-GDPR nor icons related.

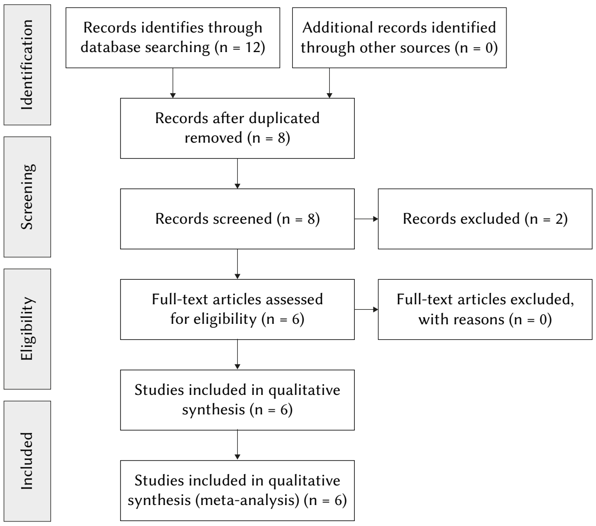

Identification
Screening
Eligibility
Included
Records identifies through database searching (n = 12)
Additional records identified through other sources (n = 0)
Records after duplicated removed (n = 8)
Records screened (n = 8)
Records excluded (n = 2)
Full-text articles assessed for eligibility (n = 6)
Full-text articles excluded, with reasons (n = 0)
Studies included in qualitative synthesis (n = 6)
Studies included in qualitative synthesis (meta-analysis) (n = 6)

Fig. 1. PRISMA’s flow diagram of the SLR conducted.

In response to MQ1, six related articles have been published in the last six years, one in 2017, one in 2018, one in 2019, two in 2020, and one in 2021. In response to MQ2, two articles have been published in a scientific journal and 4 have been published in scientific conferences. In response to MQ3, the authors who publish these works are represented in Table I, with Rossi and Palmirani being the most active and repeated actors in the works found:

TABLE I. Authors

Author	N° of papers found	Reference
Rossi, A.	5	[38]–[42]
Palmirani, M.	4	[38]–[40], [42]
Lenzini, G.	1	[41]
Martoni, M.	1	[40]
Hagan, M.	1	[40]
de Jong, S.	1	[43]
Spagnuelo, D.	1	[43]

Reading the six references allows answering RQ1 and RQ2 in terms of classification of icons related to the GPDR and Learning Analytics. From the scientific literature found, we extract that some authors have based their work on different proposals for iconified representations, including some before the enactment of the RGPD. Jong and Spagnuelo [43] classified icons into two main groups, dividing them into many subcategories:

- Data collection: personal data, sensitive data, sharing with third parties, data security, and data retention.
- Processing purposes: privacy settings, policy changes, legal obligations, user tracking, and profiling.

JJong and Spagnuelo are not the only ones to create icon taxonomies related to data privacy. Rosi and Palmirani [38]–[42] also classified icons but in multiple categories, being their taxonomy the most complete and exhaustive work among all the search results, which even considers proposals before the enactment of GDPR:

- Types of data: processed data, inferred data, etc.
- Functions of the agents: owner of the data, data controller, etc.
- Processing operations: copy, transfer of data outside the EU, etc.
- Rights of the interested parties: the right of deletion, the right of rectification, etc.
- Processing purposes: statistical purposes, research purposes, security purposes, service provision purposes, and service improvement purposes.
- Legal bases: consent, contract, legal obligation, etc.

Despite the above, not all authors equally agree about using icons to complement the information for data subjects. Institutions that use icons or other nuances to facilitate self-determined choices must consider associated risks. The data subjects must accept the terms and conditions of the services once they are fully informed; therefore, consent must be informed. However, as Efroni et al. [44] state, "the process of giving consent is often uninformed and does not encourage self-determination"; and continues that "one of the key reasons for the lack of informed consent is that users do not adequately assess or even recognize the risks (or the possible negative consequences) involved in the treatment of your data". The risk-based approach in the design of "privacy icons" as stated by Efroni et al. must consider both individual and societal risks. Consequently, there is a disparity of opinions regarding whether the icons are complementary tools enough to allow a self-determined choice before consent, or additional ones are required to fulfill this task.

Despite the legal connotations of the GDPR regarding the use of icons and the efforts made by different authors, the current proposals for the representation of Articles 13 and 14 in icons (Privacy Icons as established by Efroni et al.) do not represent some of the "processing purposes" of Learning Analytics. Rossi and Palmirani define a subcategory inside "processing purposes" as "statistical purposes". However, "statistical purposes" is a category too broad to fully inform the data subject about the treatment detail of his or her data in Learning Analytics processes. A Learning Analytics process in education refers, and not only, to the analytical treatment of student data, where different techniques and methods can be used, such as:

- Predictive Analytics
- Descriptive Analytics
- Diagnostic Analytics
- Prescriptive Analytics
- Machine Learning
- Deep Learning
- Big Data
- Artificial Intelligence
- Neural Networks
- ...

These actions are only those related to data analysis. However, there are other related actions such as the use of cookies, the internal transfer of data between departments, the storage of data for a certain time, or the processing of data for a certain period which, in part, are already included in the studies found, but that in Learning Analytics require adaptations beyond the analytical purpose, due to their connotations of fragility and sensitive context.

We confirm with this literature review that the icons designed to date include the generalities identified in the GDPR. However, they do not reach all the actions derived from using Learning Analytics. For both legal and ethical reasons, it is necessary to expand the scope of the icons with new designs to provide the maximum amount and accuracy of information to students about data processing in specific situations and contexts of Learning Analytics.

B. Phase 2

The project aims to provide students with icons to 1) generate awareness and knowledge about data processing by academic institutions and 2) make decisions based on accurate information. As stated in the introduction, this awareness and decision-making are possible if visual and standardized information is delivered clear, quick to understand, and intuitive. Considering the latter, we establish a fundamental requirement for the design of standardized icons between the execution of the study: users must know what the icon intends to inform them with only their observation. After the subsequent results of the analysis, we found that fulfilling this task depends on the subjectivities and beliefs of the participants, thus making standardization difficult and almost an impossible task.

We follow a qualitative-quantitative methodology based on conducting surveys to achieve the objectives of this phase. These surveys are executed in an iterative process, considering each iteration as a stage that depends on the previous one.

1. Materials

The materials used to develop the methodology of this second phase are mainly research instruments based on questionnaires. These questionnaires aim to collect the participants' perceptions regarding a series of icons associated with actions within Learning Analytics processes. Knowing the perception of the surveyed participants allows us to accept or discard icon designs considering consensus among responses.

Considering the iterative methodology, at each stage, one or more of these actions are performed:

- Survey the participants using multiple, open-ended, or drawing questions.
- Modify or create icons after analyzing the survey results.
- Create a new questionnaire from the new icons.

In the first iteration, we start with icons designed considering the results of the literature review carried out in phase 1. At the end of each stage of phase 2 and after analyzing the results, we create new icons or adapt those presented in the questionnaires. The questionnaires are made up of questions such as:

- Multiple answers. The question presents an icon, and participants must choose between five answers.
- Open-text answer. The question presents an icon, and participants must explain what the icon represents. Participants can indicate what modifications should be applied to make the icon clearer.
- Drawing response. The question shows an icon description, and participants must draw its graphic representation.

The resulting icons are presented in the following iterative stage. This cyclical methodology allows making changes justified by the participants' perceptions in a user experience loop.

2. Participants

The population is considered representative as both teachers and students are surveyed. Both are mainly interested in the use and visualization of the icons. A total of 103 people make up the surveyed population. Considering numbers, 15 are professor-researchers, and 88 are students. These amounts exceed the studies by Rossi and Palmirani, whose population sample is approximately 30 participants.

The project so far is divided into nine iterative stages:

- Stage 1: Creation of the first sketches. The population surveyed: 15 professor-researchers from the authors' research groups and 2 university students.
- Stage 2: Changes in some of the designs of the first phase and the addition of new ones. The surveyed population: 2 university students.
- Stage 3: Questionnaire fulfilling to validate the icons created in the previous stage. The surveyed population: 10 students.
- Stage 4: Modification of the icons considering the results of the previous stage. Conduction of questionnaires to validate the modifications. The surveyed population: 12 students.
- Stage 5: Conducted a questionnaire in which students must draw

icons based on a brief description. The objective is to acquire new design perspectives from the participants on some icons that pose problems of interpretation. The surveyed population: 15 students.

- Stage 6: Creation and modification of the icons obtained in the two previous phases.
- Stage 7: Questionnaire fulfilling to validate the icons created in the previous stage. The surveyed population: 30 students.
- Stage 8: Creation and modification of the icons obtained in the previous stage. Completion of a questionnaire to validate the newly designed icons. The surveyed population: 19 students.
- Stage 9: Last modification of the icons obtained in the previous phase (pending validation).

C. Results

Each stage of the methodology yields a series of icons validated by questionnaires. We expose some of those resulting icons for every stage. At the end of this section, we show a summary table with all the last icons resulting from the work done.

1. Stage 1

Fig. 2 exposes the first icons designed for validation in stage 1. In this stage, the very first icons are validated. The data transfer icon has 100% of consensus among the surveyed participants, not being as this regarding the icon representing the analytical treatment of data. The icons about predictive and prescriptive analytics icons needed iterative design changes to get interpretable.

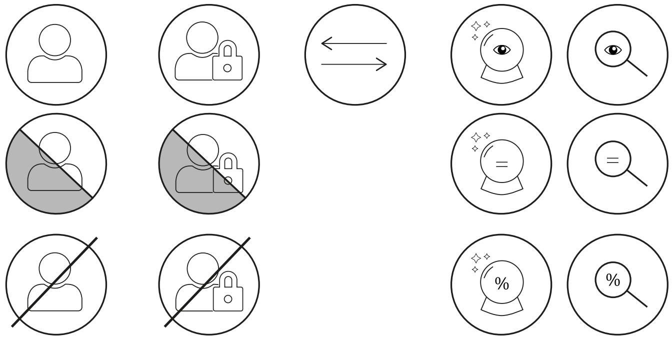


Fig. 2. First icons represent the collection of open and anonymized user data, no data collection, data transfer of any kind, and two variants of descriptive, prescriptive, and predictive data analysis using a magnifying glass and a crystal ball.

2. Stage 2

Fig. 3 shows some of the icons' modifications in stage 1 after surveying the participants. In this stage, the shape that identify a user is filled in black, and the document icon begins to be used to design icons regarding data collection.

3. Stage 3

Fig. 4 shows the designs elaborated after analyzing the results of the stage 2 survey. Those icons regarding data storage are designed considering a period. Numbers are added to the icons to identify the duration of data stored or treated.

4. Stage 4

Fig. 5 shows some of the (re)designed icons considering the results of the stage 3 survey. Changes to icons regarding analytical processing of data are made, where a magnifying glass is used instead of a crystal ball.

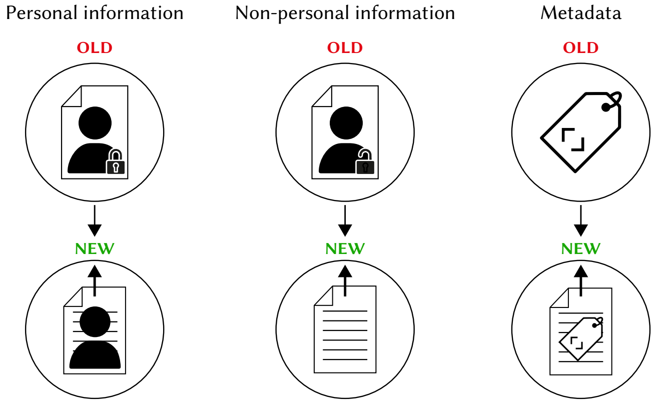


Fig. 3. Icons that identify the collection of "personal information", "non-personal information", and "metadata".

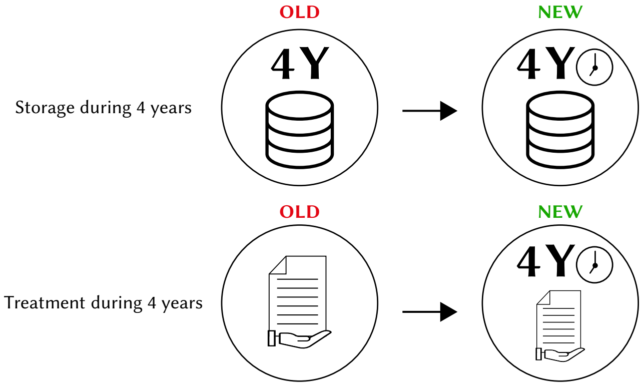


Fig. 4. Icons that identify, in this order, "data storage for 4 years" and "data treatment for 4 years".

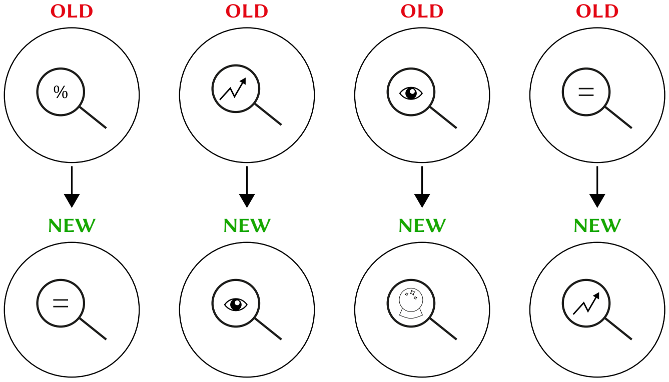


Fig. 5. Icons that identify, in this order, "descriptive analytics", "diagnostic analytics", "predictive analytics", and "prescriptive analytics".

5. Stage 5

Due to the inconclusive results, we ask the participants to draw a graphic representation of the icons. Fig. 6, Fig. 7, and Fig. 8 show some examples of the drawings made by the participants. Depending on the icon the results are more consistent, but others have pronounced differences.

6. Stage 6

Fig. 9 and Fig. 10 present the evolution of some icons between stages 5 and 6. Magnifying glasses are forgotten, and new icons are designed based on a collage. The collage idea is considered after analyzing the icons drawn by users.

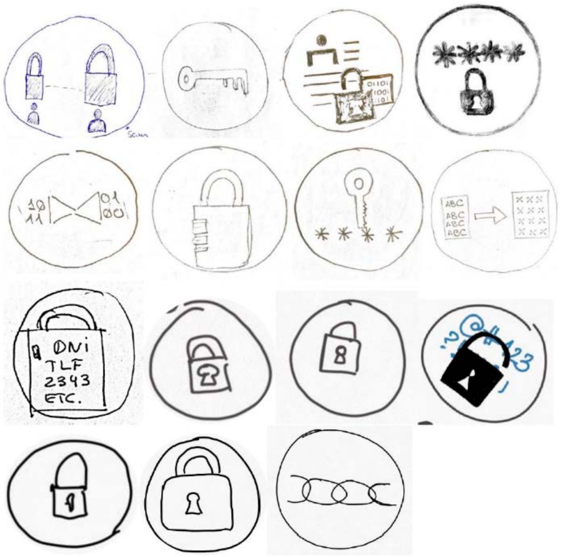


Fig. 6. Icons that identify “data encryption” drawn by participants.

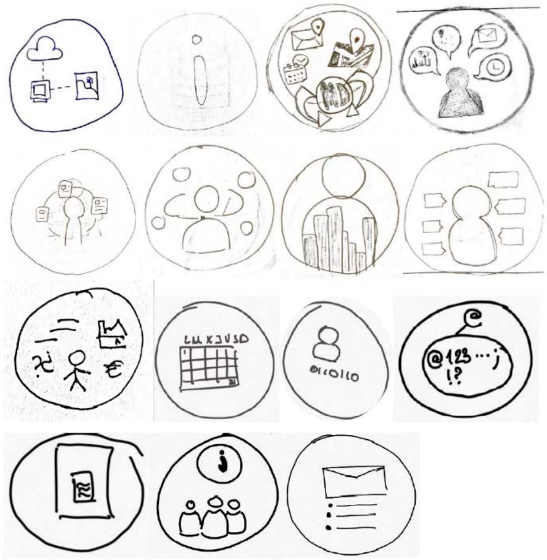


Fig. 7. Icons that identify “metadata collection” drawn by participants.

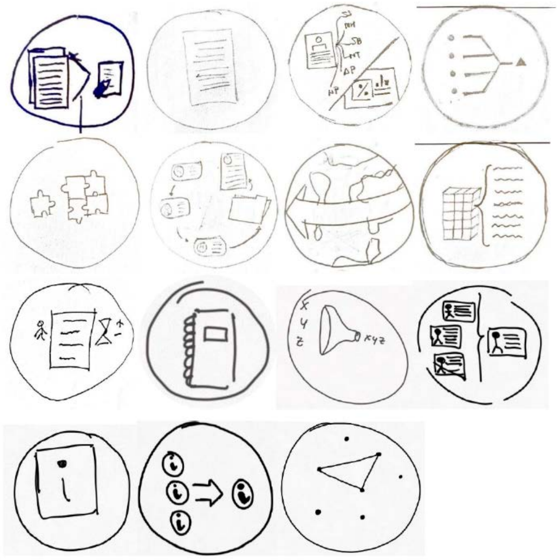


Fig. 8. Icons that identify “descriptive analytics” drawn by participants.

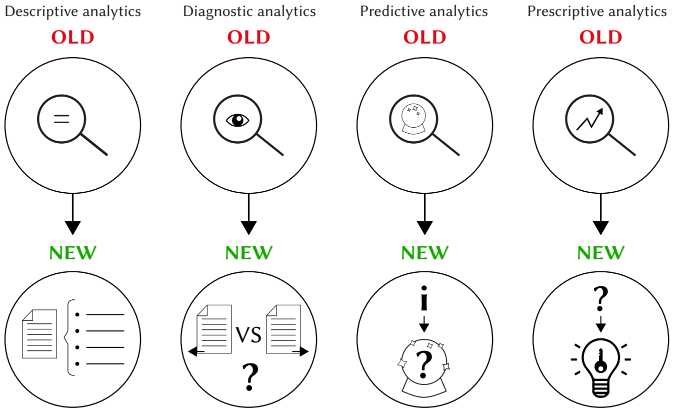


Fig. 9. Icons that identify, in this order, “descriptive analytics, “diagnostic analytics”, “predictive analytics”, and “prescriptive analytics”.

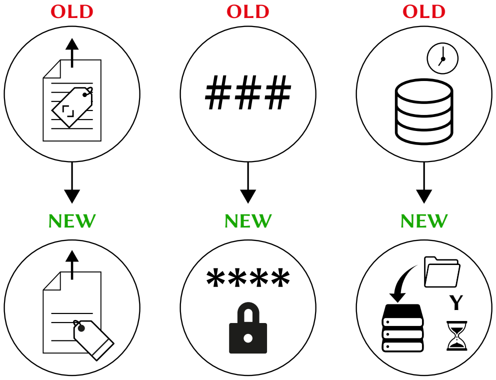


Fig. 10. Icons that identify the collection of “metadata collection”, “data encryption”, and “data storage for years”.

7. Stage 7

Fig. 11 presents the evolution of some icons between stages 6 and 7. The idea of a crystal ball returns and icons are redesigned. The shape that identifies a document is used as the base for the icons that represent data analysis.

8. Stage 7

Fig. 11 presents the evolution of some icons between stages 6 and 7. The idea of a crystal ball returns and icons are redesigned. The shape that identifies a document is used as the base for the icons that represent data analysis.

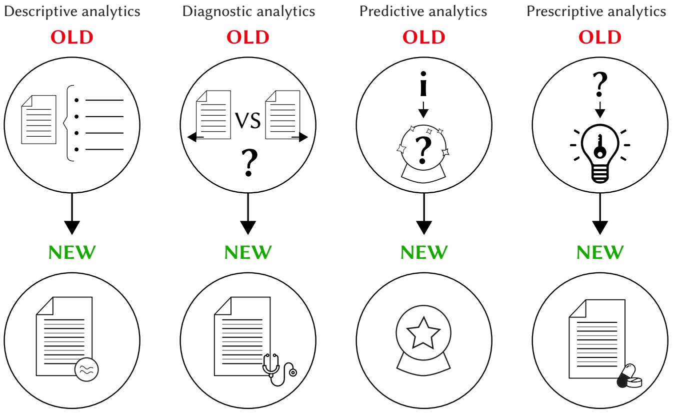


Fig. 11. Icons that identify, in this order, “descriptive analytics, “diagnostic analytics”, “predictive analytics”, and “prescriptive analytics”.

9. Stage 8

Fig. 12 and Fig. 13 present the evolution of some icons between stages 7 and 8. In general, the use of a shape that identifies a document facilitates the comprehension of the icon regarding data collection or treatment. The icon identifying the collection of metadata needs to be redesigned and the tag shape is used to identify "other data" of users.

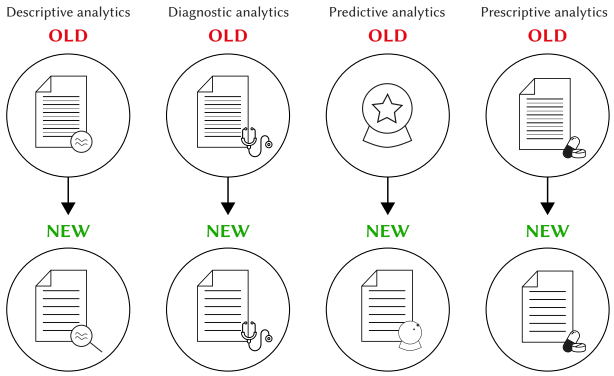


Fig. 12. Icons that identify the evolution of "descriptive analytics, "diagnostic analytics", "predictive analytics", and "prescriptive analytics".

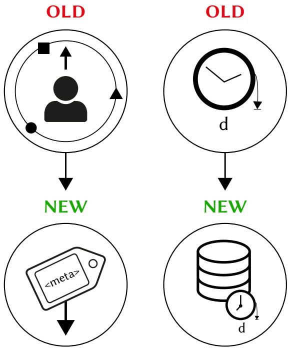


Fig. 13. Icons that identify, in this order, "metadata collection" and "data storage for days".

10.Stage 9

In this stage, we show resulting icons throughout the execution of the project using categories proposed by Jong and Spagnuelo, and Rosi and Palmirani (data collection, data storage, and processing operations). However, this is an approach that we consider not definitive and in which we are working to receive more feedback from students. For instance, there is no consensus in icons regarding "predictive analytics", where Fig. 14 shows both proposals being validated. The same happens with "cache technics", where the type of graphics inside icons seems to generate divergence, as shown in Fig. 15.

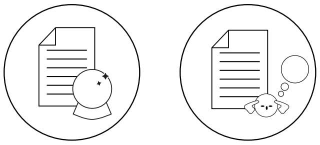


Fig. 14. Icons that identify "predictive analytics".

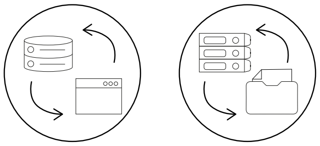


Fig. 15. Icons that identify "cache technics".

The icons extracted so far related to Learning Analytics processes are available in Table II and Table III, where its actions representation regarding data collection, data storage, and processing operations are exposed:

- Data is encrypted. Data encryption during data collection or storage.
- Data is anonymized. Anonymization of data during data collection or storage.
- Data is pseudonymized. Pseudonymization of data during data collection or storage.
- Personal information. Data collected or stored can identify data subjects.
- Metadata. Metadata collected or stored where metadata could be any non-personal data.
- Cookies. Use of web browser cookies.
- Cache technics. Use any cache technic in user devices (such as browser database), servers, or cloud computing.
- Storage for second(s), minute(s), hour(s), day(s), month(s), or year(s). Data storage of any kind for an estimated period.
- Descriptive analytics. Automated or manual descriptive data analytics approaches.
- Diagnostic analytics. Automated or manual diagnostic data analytics approaches.
- Predictive analytics. Automated or manual predictive data analytics approaches.
- Prescriptive analytics. Automated or manual prescriptive data analytics approaches.
- Data transfer to third parties. Data transfer outside the institution both in the European space or another accepted country outside European space.
- Internal data transfer. Data transfer inside the academic institution, such as between departments.
- Data treatment for second(s), minute(s), hour(s), day(s), month(s), or year(s). Data treatment of any kind for an estimated period. It differs from data storage, due data can be stored longer than treated, or vice versa.

For the "data collection" and "data storage" categories, 14 icons have been accepted by participants and are shown in Table II.

For "processing operations", 12 icons have been designed and validated by participants and shown in Table III.

D. Discussion

Designing, validating, and adapting icons that report data processing in Learning Analytics processes is an arduous task. After eight iterations, there is still a long way to go. The results obtained are very encouraging, and the icons achieved can be almost considered definitive. However, resulting icons will be subject to change as the laws and the GDPR are constantly evolving. Thus, this project will continue iterating icon designs to adapt to future law considerations.

After the collection and analysis of the results throughout the eight iterations, we affirm that:

1. It is very complex to create an icon that the affected subject knows what it is referring to just by viewing it. In this sense, it is necessary to accompany the icons with descriptive text with all the details to facilitate the comprehension of how the data will be treated.
2. It is possible to create icons that identify specific tasks and types of data analysis. Despite the difficulties expressed in the previous point, some icons show a 100% agreement among population responses; in other cases, the population agrees in 80%-95%. These

TABLE II. Icons Relative to Data Collection & Storage

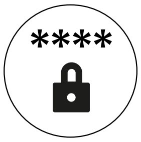
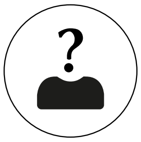
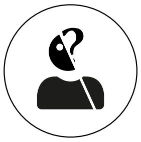
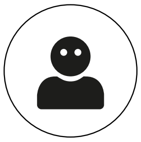
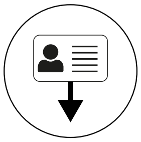
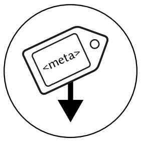
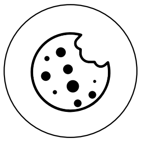
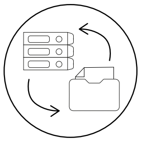
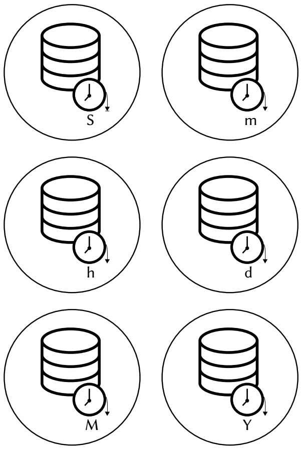
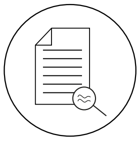
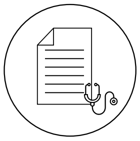
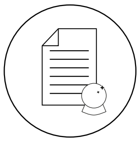
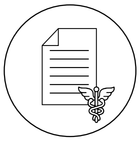
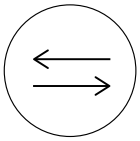
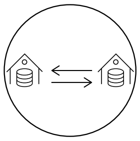
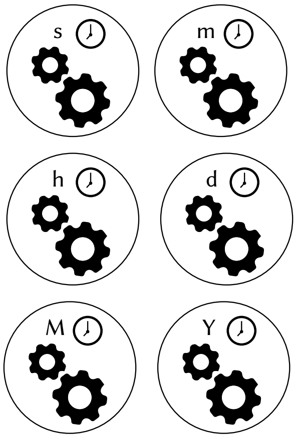
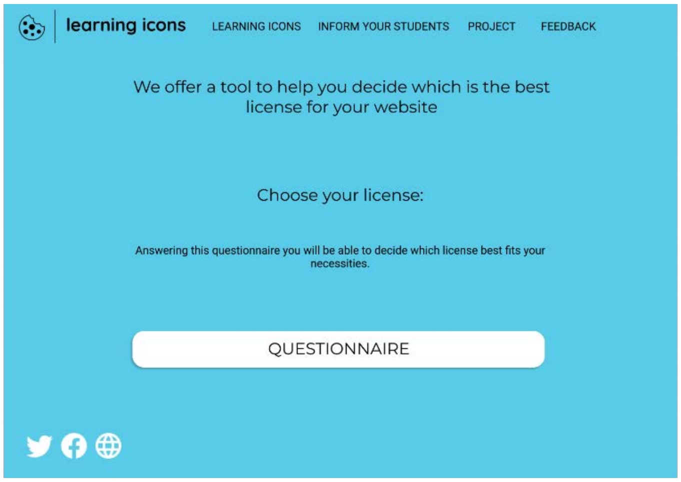
Icon description	Icon image
Data is encrypted	
Data is anonymized	
Data is pseudonymized	
Data is non-anonymized	
Personal information	
Metadata	
Cookies	
Cache technics	
Storage for second(s), minute(s), hour(s), day(s), month(s), or year(s)	

TABLE III. Icons Relative to Processing Operations

Icon description	Icon image
Descriptive analytics	
Diagnostic analytics	
Predictive analytics	
Prescriptive analytics	
Data transfer to third parties	
Internal data transfer	
Data treatment for second(s), minute(s), hour(s), day(s), month(s), or year(s)	

results indicate that the affected subjects can extract the general purpose of the icon, generating enough interest to end up reading the informative texts that accompany them if detail is needed.

We have found limitations throughout the execution of the project. The most important has been the COVID-19 pandemic. This project started in the 2019-2020 academic year. However, the COVID-19 pandemic stopped the project until we restarted it in the 2020-2021 academic year. Another limitation is the data subject's perception regarding legal terms, data processing, and Learning Analytics. These subjective perceptions lengthened the execution period of the project since there was not much consensus on the answers in the first icons iterations. Until the sixth iteration, the

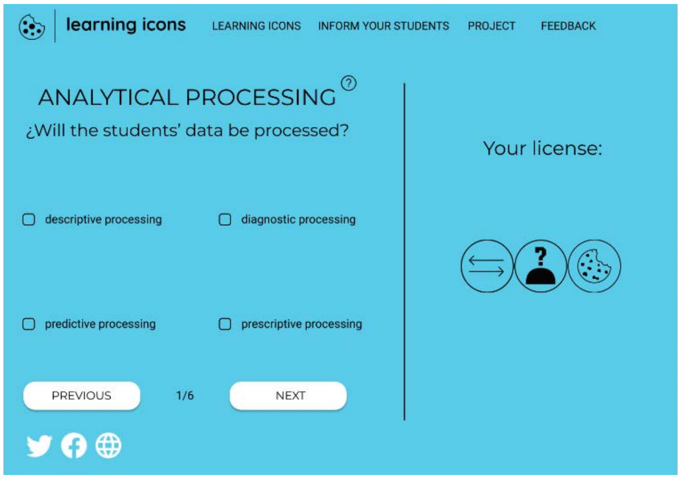


We offer a tool to help you decide which is the best license for your website

Choose your license:

Answering this questionnaire you will be able to decide which license best fits your necessities.

Fig. 16. The platform facilitates the creation of grouped icons with a questionnaire.



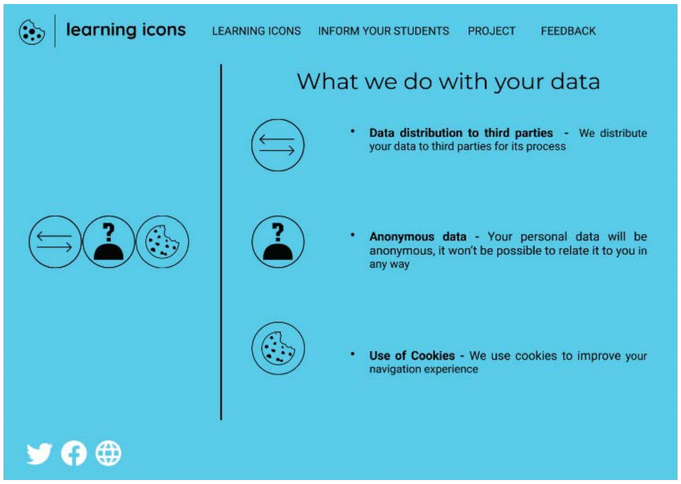
ANALYTICAL PROCESSING

¿Will the students' data be processed?

□ descriptive processing
□ diagnostic processing
□ predictive processing
□ prescriptive processing

Your license:

Fig. 17. The platform allows choosing what data treatment will be conducted.



What we do with your data

- **Data distribution to third parties** - We distribute your data to third parties for its process
- **Anonymous data** - Your personal data will be anonymous, it won't be possible to relate it to you in any way
- **Use of Cookies** - We use cookies to improve your navigation experience

Fig. 18. The platform allows showing extended text for data treatment details.

redesigns were focused on creating as much consensus as possible. Afterward, we focused on enhancing the details of the icons to improve the accuracy of meaning.

The delay in achieving the first objective implies that we could partially complete the second objective. We have begun to develop and move towards its achievement. In this sense, we present in Fig. 16, Fig. 17, and Fig. 18 the screens of the web application as a platform to facilitate anyone to easily create icons regarding data treatment of students in Learning Analytics processes. The platform allows icons to be grouped into a single image linked to an informative space displaying its meaning regarding the educational context of data processing. Considering the Article 12.7 of the GDPR, each image will be accompanied by a JSON file so icons can be machine-readable.

E. Conclusion

The GDPR provides the possibility to accompany the information provided to the data subjects identified in articles 12, 13, and 14 and recital 60 with icons that can "provide in an easily visible, intelligible and legible manner, a meaningful overview of the intended processing" [34]. Different works before the enactment of GDPR and some after identifying categories and subcategories related to data collection, purposes of the processing, types of data, functions of agents, processing operations, rights of data subjects, purposes of the processing, the legal bases.

The icons resulting from these works identify different aspects of data processing indicated in the GDPR. However, they do not cover the full range of possibilities in other contexts where data is processed constantly in different manners. This is the case in the educational context, where analytical processes such as Learning Analytics are applied. Learning Analytics is an approach in which confidential student data is processed, generating profiles and aggregated data. According to the GDPR, students must be well informed, even before enrolling in any academic course.

Our work aims to generate a series of icons, which, complemented with the work of other authors, cover specific aspects of Learning Analytics. In this way, we hope to facilitate the understanding of the treatment of student data, considering the peculiarities of each educational institution. In the manuscript, we present a two-folded methodology. On the one hand, a documental methodology based on a systematic literature review with very limited results. These very limited results indicate that there is an open field for research, especially when regulations in Europe are susceptible to recurrent changes in terms of privacy. On the other hand, the qualitative-quantitative methodology used in the design of icons for the information students' data treatment in Learning Analytics processes. Participants are part of the academic field, in specific teachers and students; we used questionnaires to collect participants' perceptions to validate icons. We present all the icons generated in the results of the applied methodology. However, the surveyed population is Spanish, and in the following iterations, we will consider participants from other European countries.

As a second objective, we set the development of a platform to facilitate the integration of icons in any VLE. This platform will allow the creation of grouped informative icons linked to an explanatory text. This text is personalized considering the context of data processing -descriptive analyzes can be conducted in one course and predictive analyzes in another-. The platform is currently in development. Thus, the second objective is not accomplished. However, we present screens to appreciate some characteristics and functionalities.

Future work focuses on two phases. First, finish the Learning Analytics Icons platform. Second, extend participation from other countries. Subjectivities of participants are associated with cultural implications. Icons designed should be tested in countries different from Spain to create standard icons or adapted ones to the countries if participants' consensus is very divergent.

Funding

This research was funded by EU within the framework of the Erasmus+ Programme. Project Security and Privacy in Academic

Data management at Schools (SPADATAS) Grant Agreement n° 2022-1-ES01-KA220-000086363.

Acknowledgment

Many thanks to Aleix Ollé, Silvia Blasi, Javier Geli, and Rogelio Sansaloni who have helped in undertaking the research work, specially in the execution of surveys and icon design.

References

[1] C. Romero, S. Ventura, M. Pechenizkiy, and R. S. J. D. Baker, *Handbook of educational data mining*. CRC press, 2010, p. 503. doi: 10.1201/b10274.

[2] M. A. Chatti, A. L. Dyckhoff, U. Schroeder, and H. Thüs, "A reference model for learning analytics," *International Journal of Technology Enhanced Learning*, vol. 4, no. 5–6, pp. 318–331, 2012, doi: 10.1504/IJTEL.2012.051815.

[3] F. J. García-Peñalvo, "Learning Analytics as a Breakthrough in Educational Improvement," in *Radical Solutions and Learning Analytics*, Springer, 2020, pp. 1–15. doi: 10.1007/978-981-15-4526-9_1.

[4] D. Amo, M. Alier, M. J. Casan, and M. J. Casañ, "The student's progress snapshot a hybrid text and visual learning analytics dashboard," *The International Journal of Engineering Education*, vol. 34, no. 3, pp. 990–1000, 2018.

[5] G. Siemens, "What are Learning Analytics?," *Elearnspace*, no. 1, pp. 1–1, 2010. Accessed: Jan. 01, 2015. [Online]. Available: https://web.archive.org/web/20100827114932/http://www.elearnspace.org/blog/2010/08/25/what-are-learning-analytics/

[6] D. Amo *et al.*, "Using Web Analytics Tools To Improve the Quality of Educational Resources and the Learning Process of Students in a Gamified Situation," in *INTED2018 Proceedings*, 2018, vol. 1, pp. 5824–5829. doi: 10.21125/inted.2018.1384.

[7] M. Alier, M. J. Casany, C. Severance, and D. Amo, "Learner Privacy, a pending assignment," in *Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*, New York, NY, USA, Oct. 2020, pp. 725–729. doi: 10.1145/3434780.3436635.

[8] M. G. Alonso de Castro and F. J. García-Peñalvo, "Successful educational methodologies: Erasmus+ projects related to e-learning or ICT," *Campus Virtuales*, vol. 11, no. 1, pp. 95–114, 2022, doi: 10.54988/cv.2022.1.1022.

[9] D. Amo, "MOOCs: Experimental approaches for quality in pedagogical and design fundamentals," in *ACM International Conference Proceeding Series*, 2013, pp. 219–223. doi: 10.1145/2536536.2536570.

[10] Z. N. Khlaif, M. Ghanim, A. A. Obaid, S. Salha, and S. Affouneh, "The Motives and Challenges of developing and delivering MOOCs courses," *Education in the Knowledge Society*, vol. 22, p. art. e23904, 2021, doi: 10.14201/eks.2390.

[11] M. Á. Conde-González, F. J. García-Peñalvo, M. J. Rodríguez-Conde, M. Alier, and A. García-Holgado, "Perceived openness of Learning Management Systems by students and teachers in education and technology courses," *Computers in Human Behavior*, vol. 31, pp. 517–526, 2014, doi: 10.1016/j.chb.2013.05.023.

[12] Á. Fidalgo-Blanco, M. L. Sein-Echaluce, F. J. García-Peñalvo, and M. Á. Conde, "Using Learning Analytics to improve teamwork assessment," *Computers in Human Behavior*, vol. 47, pp. 149–156, 2015, doi: 10.1016/j.chb.2014.11.050.

[13] R. F. Kizilcec, C. Piech, and E. Schneider, "Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses," in *ACM International Conference Proceeding Series*, 2013, pp. 170–179. doi: 10.1145/2460296.2460330.

[14] X. Ochoa, N. Weibel, M. Worsley, and S. Oviatt, "Multimodal learning analytics data challenges," in *6th International Conference on Learning Analytics and Knowledge, LAK 2016*, 2016, pp. 498–499. [Online]. Available: https://sci-hub.se/https://nyu-staging.pure.elsevier.com/en/publications/multimodal-learning-analytics-data-challenges

[15] A. Álvarez-Arana, M. Villamañe-Gironés, and M. Larrañaga-Olagaray, "Improving Assessment Using Visual Learning Analytics," *Education in the Knowledge Society*, vol. 21, no. 9, pp. 1–9, 2020, doi: 10.14201/eks.21554.

[16] D. Amo, R. Torres, X. Canaleta, J. Herrero-Martín, C. Rodríguez-Merino, and D. Fonseca, "Seven principles to foster privacy and security in educational tools: Local Educational Data Analytics," in *TEEM'20: Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*, 2020, pp. 730–737. doi: 10.1145/3434780.3436637.

[17] H. Drachsler and W. Greller, "Privacy and analytics: it's a DELICATE issue a checklist for trusted learning analytics," in *Proceedings of the sixth international conference on learning analytics & knowledge*, Edinburgh, United Kingdom, 2016, pp. 89–98. doi: 10.1145/2883851.2883893.

[18] F. J. García-Peñalvo and A. Corell, "La COVID-19: ¿enzima de transformación digital de la docencia o reflejo de una crisis metodológica y competencial en la educación superior?," *Campus Virtuales*, vol. 9, no. 2, pp. 83–98, 2020.

[19] F. J. García-Peñalvo, A. Corell, R. Rivero-Ortega, M. J. Rodríguez-Conde, and N. Rodríguez-García, "Impact of the COVID-19 on Higher Education: An Experience-Based Approach," in *Information Technology Trends for a Global and Interdisciplinary Research Community*, IGI Global, 2021, pp. 1–18. doi: 10.4018/978-1-7998-4156-2.ch001.

[20] F. J. García-Peñalvo, A. Corell, V. Abella-García, and M. Grande, "Online assessment in higher education in the time of COVID-19," *Education in the Knowledge Society*, vol. 21, pp. 12–26, 2020, doi: 10.14201/eks.23013.

[21] D. Amo, M. Alier, F. J. García-Peñalvo, D. Fonseca, and M. J. Casany, "GDPR security and confidentiality compliance in LMS' a problem analysis and engineering solution proposal," in *TEEM'19: Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*, León, 2019, pp. 253–259. doi: 10.1145/3362789.3362823.

[22] Y. Jang, R. Katz, and K. Dalkir, "Are Higher-Education Institutions Ready for Learning Analytics? Governance, Ethics, Confidentiality and Privacy," *Journal of Leadership, Accountability and Ethics*, vol. 18, no. 1, pp. 137–148, 2021.

[23] D. Amo, M. Alier, F. García-Peñalvo, D. Fonseca, and M. J. Casañ, "Privacidad, seguridad y legalidad en soluciones educativas basadas en Blockchain: Una Revisión Sistemática de la Literatura," *RIED. Revista Iberoamericana de Educación a Distancia*, vol. 23, no. 2, pp. 213–236, 2020, doi: 10.5944/ried.23.2.26388.

[24] J. Seanosky, D. Jacques, and V. Kumar, "Security and Privacy in Bigdata Learning Analytics," in *Proceedings of the 3rd International Symposium on Big Data and Cloud Computing Challenges (ISBCC–16')*, Cham, Switzerland, 2016, pp. 43–55. doi: 10.1007/978-3-319-30348-2_4.

[25] F. Jose Garcia-Penalvo, "Digital Identity as Researchers. The Evidence and Transparency of Scientific Production," *Education in the Knowledge Society*, vol. 19, no. 2, pp. 7–28, 2018, doi: 10.14201/eks201819272.

[26] M. Alier, M. J. Casany, C. Severance, and D. Amo, "Learner Privacy, a pending assignment," in *ACM International Conference Proceeding Series*, 2020, pp. 725–729. doi: 10.1145/3434780.3436635.

[27] B. Herold, "InBloom to Shut Down Amid Growing Data-Privacy Concerns," *Education Week*, 2014. [Online]. Available: http://blogs.edweek.org/edweek/DigitalEducation/2014/04/inbloom_to_shut_down_amid_growing_data_privacy_concerns.html

[28] M. Grothaus, "Pearson data breach: details of hundreds of thousands of U.S. students hacked," *Fast Company*, 2019. [Online]. Available: https://www.fastcompany.com/90384759/pearson-data-breach-details-of-hundreds-of-thousands-of-u-s-students-hacked

[29] L. Amoore, "Why 'Ditch the algorithm' is the future of political protest," *The Guardian*, 2020. [Online]. Available: https://www.theguardian.com/commentisfree/2020/aug/19/ditch-the-algorithm-generation-students-a-levels-politics

[30] A. Pardo and G. Siemens, "Ethical and privacy principles for learning analytics," *British Journal of Educational Technology*, vol. 45, no. 3, pp. 438–450, May 2014, doi: 10.1111/bjet.12152.

[31] S. Slade and P. Prinsloo, "Learning Analytics: Ethical Issues and Dilemmas," *American Behavioral Scientist*, vol. 57, no. 10, pp. 1510–1529, 2013, doi: 10.1177/0002764213479366.

[32] F. J. García-Peñalvo, "Avoiding the dark side of digital transformation in teaching. An institutional reference framework for eLearning in higher education," *Sustainability*, vol. 13, no. 4, p. art. 2023, 2021, doi: 10.3390/su13042023.

[33] EP and the CEU, "Regulation (EU) 2016/679 GDPR," *Official Journal of the European Union*, pp. 88–88, 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32016R0679

[34] E. and the CEU, "Recital 60 - Information obligation | General Data Protection Regulation (GDPR)." [Online]. Available: https://gdpr-info.eu/

recitals/no-60/

[35] A. García Holgado, S. Marcos Pablos, and F. J. García Peñalvo, “Guidelines for performing systematic research projects reviews,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 2, pp. 136–144, 2020, doi: 10.9781/ijimai.2020.05.005.

[36] F. J. García-Peñalvo, “Developing robust state-of-the-art reports: Systematic Literature Reviews,” *Education in the Knowledge Society*, vol. 23, 2022.

[37] D. Moher *et al.*, “Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement,” *PLoS Medicine*, vol. 6, no. 7, pp. e1000097–e1000097, Jul. 2009, doi: 10.1371/journal.pmed.1000097.

[38] A. Rossi and M. Palmirani, “DaPIS: An Ontology-Based Data Protection Icon Set,” *Knowledge of the Law in the Big Data Age*, vol. 317, no. 978-1-61499-984–3, pp. 181–195, 2019, doi: 10.3233/FAIA190020.

[39] A. Rossi and M. Palmirani, “Can visual design provide legal transparency? The challenges for successful implementation of icons for data protection,” *Design Issues*, vol. 36, no. 3, pp. 82–96, 2020, doi: 10.1162/desi_a_00605.

[40] M. Palmirani, A. Rossi, M. Martoni, and M. Hagan, “A methodological framework to design a machine-readable privacy icon set,” in *Jusletter IT*, 2018, no. February.

[41] A. Rossi and G. Lenzini, “Which Properties Has an Icon? A Critical Discussion on Data Protection Iconography,” in *Socio-Technical Aspects in Security and Trust*, Cham, 2021, pp. 211–229. doi: 10.1007/978-3-030-55958-8_12.

[42] A. Rossi and M. Palmirani, “A Visualization Approach for Adaptive Consent in the European Data Protection Framework,” in *2017 Conference for E-Democracy and Open Government (CeDEM)*, May 2017, pp. 159–170. doi: 10.1109/CeDEM.2017.23.

[43] S. de Jong and D. Spagnuelo, “Iconified representations of privacy policies: A GDPR perspective,” in *Advances in Intelligent Systems and Computing*, 2020, vol. 1160, pp. 796–806. doi: 10.1007/978-3-030-45691-7_75.

[44] Z. Efroni, J. Metzger, L. Mischau, and M. Schirmbeck, “Privacy icons: A risk-based approach to visualisation of data processing,” *European Data Protection Law Review*, vol. 5, no. 3, pp. 352–366, 2019, doi: 10.21552/edpl/2019/3/9.

Daniel Amo-Filva

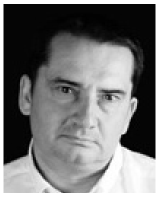
Ph.D. in Education Sciences from the University of Salamanca (2020), with two master’s degrees in education and educational technology, a University Master’s Degree in Teacher Training for Compulsory Secondary Education, and a Baccalaureate, Professional Training and Language Teaching (UNIR 2016) and University Master’s Degree in Education and ICT, specialization in Research (UOC 2014). With 15 years of participating in and leading different technological projects in the tech-industry, he currently focuses his professional career on university teaching in the Department of Computer Engineering at La Salle, Ramon Lull University, and research group HER (Human-Environment Research). Within HER he coordinates the Technology-Enhanced Learning (TEL) research line focused on the design, implementation, evaluation, and improvement of the impact of educational projects in any academic fields mediated by technology. The use of ethical and analytical methods for educational data treatment are the fundamentals to the evaluation and improvement of such procedures. He actively participates in scientific congress committees and conferences to disseminate to society the knowledge resulting from his professional and personal research. He is the author of the two books “Learning Analytics: the narrative of learning through data” (UOC OuterEdu) and “Learning analytics: 30 experiences in the classroom with data”; the blog eduliticas.com where he disseminates about Learning Analytics; the podcast connecta.danielamo.info where tackles technology, privacy, and humanistic aspects of society; and danielamo.info a personal space where he shows all his past and present projects. Research publications are available at Google Scholar (https://scholar.google.com/citations?user=RNHbv9oAAAAJ&hl=es&oi=ao), and ORCID (https://orcid.org/0000-0002-4929-0438).

Marc Alier

Marc Alier (1971) received an engineering degree in computer science (1996) and a PhD in Sustainability (2009) in the Polytechnical University of Catalonia (UPC). He is an associate professor at UPC and deputy director at ICE http://www.ice.upc.edu. The last 25 years have worked in research and development related to the e-learning industry. Has participated in the development of several LMS, content authoring tools and interoperability standards. Since 2001, has taught software engineering, project management, information systems, and computing ethics at UPC’s School of Informatics. Has been director of several master’s programs. Has authored more than 120 papers in journals and conference proceedings. Since 2007 produces several podcasts about technology, science, and its impact on society as a means of dissemination of his professional and personal research.

David Fonseca

Full Professor (2017) by La Salle Ramon Llull University, currently he is the coordinator of the Group of Research on Technology Enhanced Learning (GRETEL), a recognized research group of Generalitat de Catalunya (from 2014), and coordinator of the Graphic Representation Area in the Architecture Department of La Salle (where he is a teacher and academic tutor). Technical Engineer in Telecommunications (URL – 1998), Master in GIS (Universitat de Girona, 2003), Audiovisual Communication Degree (UOC, 2006), Master in Advanced Studies (URL-2007), Official Master in Information and Knowledge Society (UOC, 2008), PhD in Multimedia by URL (2011), also, he is Autodesk Approved and Certified Instructor from 1998. With extensive experience in project manager (from 2000 to act, he has coordinated more than 50 local, national, and international projects, both technological transfer and research funded projects), he has directed 7 PhD thesis and more than 10 other final degree and master projects. Currently he is serving as program or scientific committee in more than 15 indexed journals and conferences, as well as organizing workshops, special issues and invited sessions in different scientific forums.



Francisco José García Peñalvo

Full Professor in the Department of Computer Science and Automation at the University of Salamanca (USAL), with three six-year periods of research, one six-year period of transferring and innovation, and four five-year periods of recognized teaching. He received the Gloria Begué award for teaching excellence in 2019. He was also a Distinguished Professor at the School of Humanities and Education at the Tecnológico de Monterrey, Mexico (2016-2018) and is a Researcher of International Impact at the Universidad Nacional San Agustín, Arequipa, Peru. Since 2006 he is the head of the Research Group Recognized by the USAL GRIAL (research GRoup on InterAction and eLearning), a group that is a Consolidated Research Unit of the Junta de Castilla y León Government (UIC 81). Included in the University of de Stanford’s World’s Top 2% Scientists list (2019, 2020, 2021) https://doi.org/10.17632/btchxktzyw.3. He has supervised 28 Ph.D. thesis. He has been Vice-Dean of Innovation and New Technologies of the Faculty of Sciences of the USAL between 2004 and 2007 and Vice-Rector of Technological Innovation of this University between 2007 and 2009. He is currently the Deputy Director of the Research Institute for Educational Sciences (IUCE), the Rector’s Delegate for Digital Learning and Teaching and the Coordinator of the Doctorate Programme in Education in the Knowledge Society at USAL. He is Editor-in-Chief of the journals Education in the Knowledge Society and Journal of the Information Technology Research, and Associate Editor of many journals, with a special mention to the journals IEEE Transactions on Learning Technologies, IEEE Access, Computers in Human Behavior, and Computers in Human Behavior Reports. He has published more than 100 research papers in JCR SCIE/SSCI-indexed journal (55 Q1). For more detailed information on the publications, these are the public links to the profiles in Google Scholar (http://goo.gl/sDwrr0), Publons (https://bit.ly/2u2FN5l), Scopus (https://bit.ly/3IYoog7), and ORCID (http://orcid.org/0000-0001-9987-5584).

María José Casañ

Is a computer science Engineer from Polytechnic University of Catalonia - UPC (1997) and has a Ph.D. in Science (2013) from UPC. From 2004 she has been a researcher and lecturer, teaching at the School of Informatics on UPC. She has also been a course instructor at the Open University of Catalonia UOC. She teaches courses on Software engineering Projects, Databases, Social and Environmental aspects of computing as well as history of computers. She has developed several Open source projects such as the J2MEMicroDB (a database engine for Mobile devices) and a migration of the authoring software JClic to the OLPC X0 platform (http://laptop.org). She has also participated in the development of several LMS projects and authoring tools. Her research interests focus on the social aspects of engineering education, innovation in higher education degrees, educational innovation that contributes to the quality teaching enhancement, sustainability, and techno-ethics.

Simulations for the Precise Modeling of Exercises Including Time, Grades and Number of Attempts

Alberto Jiménez-Macías*, Pedro J. Muñoz-Merino, Carlos Delgado Kloos

Universidad Carlos III de Madrid, Leganés (Spain)

* Corresponding author: albjimen@it.uc3m.es

Received 10 November 2022 | Accepted 24 May 2024 | Published 9 October 2024

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

Students' interactions with exercises can reveal interesting features that can be used to redesign or effectively use the exercises during the learning process. The precise modeling of exercises includes how grades can evolve, depending on the number of attempts and time spent on the exercises. A missing aspect is how a precise relationship among grades, number of attempts, and time spent can be inferred from student interactions with exercises using machine learning methods, and how it differs depending on different factors. In this study, we analyzed the application of different machine-learning methods for modeling different scenarios by varying the probability of answering correctly, dataset sizes, and distributions. The results show that the model converged when the probability of random guessing was low. For exercises with an average of 2 attempts, the model converged to 200 interactions. However, increasing the number of interactions beyond 200 does not affect the accuracy of the model.

Keywords

Content Modeling,
Learning Analytics,
Simulated Students,
Smart Content.

DOI: 10.9781/ijimai.2024.10.002

I. Introduction

The learning content utilized in teaching and learning is crucial because it is a valuable tool for enhancing students' understanding and influencing their cognitive and metacognitive capacities. However, its usefulness may be limited if learning content remains stagnant and cannot be expanded. Smart learning content (SLC) included advanced features, such as adaptive personalization, sophisticated feedback forms, user authentication, learner modeling, data aggregation, and learning analytics [1]. SLC can improve student engagement and success by providing personalized learning experiences that are adapted to individual needs and preferences [2].

Depending on the type of content, we can have different types of student interaction. One type of content is related to tests in which students can make different attempts to solve problems, for example, multiple responses or fill-in-the-blank exercises. Probabilistic approaches, such as the Item Response Theory (IRT), have been utilized to model educational tests. This method allows the estimation of item characteristics, such as difficulty, discrimination, and guessing, using student interactions [3]. Some authors have also used content modeling to estimate additional content parameters. For example, some models help infer the skill acquired by students after using educational materials [4].

In this work, we focus on exercises and understanding such as any activity task that should be solved by a student at some time, after many attempts, and in which the student can achieve a grade for each attempt. Examples of types of exercises include multiple choices, multiple responses, and a drag&drop, but also an open problem in which automatic evaluation is not possible, and a teacher should grade it by looking at different steps and a long text.

An exercise can be better characterized by establishing a clear relationship between the number of attempts, time spent, or grades. In the results of the systematic literature review [5], these three exercise characteristics were the next most frequently used in different studies after excluding the three used by the IRT: difficulty, discrimination, and guessing. For example, Moreno-Marcos et al. [6] used grade, time, and number of attempts to identify behavioral patterns, such as persistence, efficiency, and constancy, within an intelligent tutoring system. Also, Feng et al. [7] calculated indicators of the activities carried out by students, including the average number of attempts for each question, time spent on the activities, and number of finished activities, among others.

The term "learning curve" in the field of education pertains to the speed at which a student acquires a specific skill or set of skills. Learning curves can be used to track a learner's advancement by evaluating their performance over time, identifying areas of strength and weakness, and determining the most effective means of supporting learning [8]. In the plots of this theory, the 2D graphs relate performance (the grade obtained) to the learning effort (the number of attempts), and performance to the time employed.

Please cite this article as:
A. Jiménez-Macías, P. J. Muñoz-Merino, C. Delgado Kloos. Simulations for the Precise Modeling of Exercises Including Time, Grades and Number of Attempts, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 127-140, 2025, http://dx.doi.org/10.9781/ijimai.2024.10.002

Different studies have analyzed these three characteristics independently in models for educational exercises; however, we identified a gap when using these three characteristics simultaneously in a model and analyzed how they relate to each other. The redesign and use of exercises can be improved by understanding the relationship between these indicators. The focus of this study was to use machine learning techniques to infer the relationships among these parameters based on student interactions with exercises.

To evaluate these exercise models, we needed a significant number of interactions performed by students in those exercises. One posible solution is to use simulated students. Previous studies have used different simulated students [9], [10], [11] to recreate different student learning situations. The use of simulations in this work does not attempt to replicate the real student's behavior, but to test the models in different predefined scenarios of student behavior so that we can know, for example, the number of interactions necessary for different cases. To evaluate the accuracy of machine-learning algorithms, different metrics can be used, such as precision, recall, F1, Root Mean Square Error(RMSE) default without normalization, and Area Under the Curve(AUC) [12].

This study aimed to analyze the possibility of using machine learning methods to infer these exercise indicators using student simulations. We propose the following research questions:

- Is it possible to obtain a time-grade-attempt model in exercises that are sufficiently accurate using traditional machine learning algorithms?
- How does the accuracy of the models vary for different types of exercises?
- What is the minimum number of interactions required to stabilize an exercise model with acceptable accuracy?
- How do different forms of student behavior modify the results obtained in the previous research questions?

This paper extends our paper [13]. We analyzed a model design for educational exercises using grade, number of attempts, and time spent. We tested different machine learning algorithms using simulated data for each variable using normal distributions. The paper is structured as follows: Sections I and II of this paper (Introduction and Related Work) include ideas from [13] but extend it with new ideas and references as the research questions have been extended to analyze the effect of changing the probability of answering correctly, dataset sizes and different distributions. Subsection III-A presents an overview of the extended paper [13] and takes ideas, results, and analysis from [13] but has been rewritten to try to increase clarity; Subsection III-B includes a new analysis of the respective metrics evaluated, the same visualizations with another dataset, and a new analysis of model over-fitting, while Sections IV, V, and VI are new and analyze the behavior of this proposed model in different scenarios. Section IV presents the simulations for different types of questions, dataset sizes, and distributions, Section V shows the results obtained in the simulations, Section VI presents a discussion of the results obtained, and Section VII presents the conclusions and future work.

II. Related Work

Smart learning environments (SLEs) are learning environments capable of enhancing education by using adaptive technologies [14]. The content available to the learner and the knowledge acquired by the learner is part of this environment. Content should be constructed based on the learner's previous experience by identifying their needs and learning styles [15]. Content modeling is relevant for learning because it allows for the redesign and improvement of teachers' content, thus helping students' learning.

In content modeling, probabilistic models take advantage of content parameters to understand and represent the learning materials. These models can be evaluated through simulations or real scenarios to provide insights into their effectiveness and adaptability in intelligent learning environments. The following subsections indicate the application of probabilistic models, the importance of parameters, simulations employed, and critical consideration of the number of interactions in optimizing these models to improve educational outcomes.

A. Probabilistic Models

Among the studies in which probabilistic methods are used, the most frequently used algorithm is IRT [16] [17] for modeling items in tests. IRT can estimate exercise parameters such as difficulty, discrimination, and guessing, based on the interactions made by the students in the questionnaires. For example, IRT was used to provide individual learning paths for students, which can alleviate disorientation and cognitive overload in learners based on the difficulty of course materials and their ability to improve learning efficiency and effectiveness [18]. The authors suggest that additional research and testing is required to thoroughly assess its effectiveness and potential limitations. Abbakumov [19] used a modified version of IRT to estimate the difficulty levels of items and address the cold-start problem using an application developed at the Higher School of Economics University. Consequently, learner motivation can be maintained, frustration and stress can be reduced, and learning outcomes can be improved. The author did not indicate any limitations but promised further work to evaluate the efficacy of the proposed model using real student data and to optimize the model's performance on topics with a medium level of difficulty, which typically has regression coefficients that are relatively inconsequential.

Artificial intelligence algorithms, such as regression, random forest, and neural networks, have been used to determine the parameters of Item Response Theory (IRT) and to evaluate the accuracy of these models [20] [21] [22]. In a study [23], a regression algorithm was used to measure difficulty and discrimination in multiple-choice questions. The results were compared to those obtained using IRT to estimate the same parameters. In addition, Lehman et al. [24] analyzed the emotions that students experience during conversation-based evaluations.

Another type of content modeling has been applied to discussion forums. Capuano et al. [25] used neural networks to classify students' answers in the forums and to detect the confusion perceived by students when participating in the discussion in real-time. The suggested approach can potentially enhance interactivity and support for students in Massive Open Online Courses (MOOCs). However, the authors acknowledge that additional research is necessary to assess the effectiveness of this approach thoroughly. Neural networks were used in discussion forums to detect feelings produced by a forum for students in MOOCs [26]. The linguistic-feature-based confusion classifier performed well on the evaluated metric F1-score, allowing real-time detection of message confusion. A limitation of the study was false negatives because teachers would not be able to identify messages in need of urgent intervention.

B. Parameters

In exercises, the grade, time spent, and the number of attempts have been used in different studies as indicators, as in the work by Feng, Heffernan and Koedinger [27]. Verdú et al. [28] proposed a model based on genetic algorithms and fuzzy systems to accurately classify questions according to their difficulty level in an intelligent tutoring system. They used the following parameters in their model: time in minutes from the last reading of the question to the delivery of the answer, grade obtained for that answer, and number of accesses or readings before sending the answer.

Regarding grade, Uto [29] proposed a model to estimate the ability and grade obtained by students in written essays using the IRT model with evaluator parameters integrated into a model of the topicality of the answers. This model is based on the Latent Dirichlet Allocation(LDA) model of responses obtained by students in essays. In addition, the final grades for a subject and master's degree in a university online mode were determined using different machine learning algorithms such as Naive Bayes, Decision Tree, Random Forest, and Neural Networks [30].

In terms of time, Rushkin, Chuang, and Tingle [4] described a log-normal model to estimate the slowness of the learners and the characteristics of the evaluations, such as discrimination and time intensity, using response times in an online course. In addition, Xue, Yaneva, Runyon and Baldw [20] predicted difficulty and response time for multiple-choice questions using information from each item text in the medical examination questions.

We [13] proposed providing more details on exercises based on three characteristics: grade of each attempt, time spent in each attempt, and number of attempts using the probabilistic method. The contribution of this study is to propose a detailed analysis of how the three indicators are related using simulations. In addition, they proposed different characteristics of the most used exercises, such as difficulty, discrimination, and guessing, calculated as parameters using IRT. The results demonstrated the accuracy of the machine learning algorithms using the proposed model design, indicating that the use of simulated students was a limitation of the study.

C. Simulations

Regarding simulation in education, VanLehn et al. [31] found three main applications in which simulated students could be used: as peers of real students, in instructional pedagogical design, and teachers' learning methods. Various tools have been developed, such as SimStudent [32] and Demonstr8 [33], to test different models in a simulated environment before testing them in a real environment. These tools are helpful for the learning process because they allow the evaluation of different conditions required in the evaluated models [34].

Moreover, some systems simulate the students during the learning process. For example, Graesser [35] proposed an architecture that uses a simulation approach to implement pedagogical agents that focus on peer learning. Vizcaino [36] described an architecture in a collaborative environment that uses simulated students to detect and avoid possible scenarios that do not improve collaborative learning.

We propose the use of student simulations to recreate different possible scenarios in which the model could be used. Previous studies used simulated students to validate the models proposed by the authors. For example, Champaign and Cohen [11] proposed an approach for selecting content in an intelligent tutoring system based on student interactions. A simulated student was used to validate the proposed model and attempt to recreate a real-world scenario. However, there are clear constraints in creating simulated students that exactly match real learners. The researchers determined that their algorithm was efficient in choosing relevant educational content for students by considering the prior learning experiences of similar peers. Dorcca [9] used simulated students to evaluate three strategies in models of student learning styles, reducing the number of resources needed to validate the proposed approaches, understanding the proposed system's behavior in this scenario, and making necessary changes to improve the design. However, simulated students may not fully capture real students' behavior and responses, and the effectiveness of adaptive educational systems with simulated students may not always be generalizable to real-world settings.

D. Number of Interactions

The number of interactions or runs required in any machine-learning algorithm is important to identify the performance of any proposed model and different studies have been conducted to identify the number of interactions or runs needed. For example, Liu et al. [37] found that it was necessary to run a Bayesian Network algorithm twice. Erickson et al. [38] identified 100 interactions to determine the best approach to learning object allocation. Frost and McCalle [39] required 25 simulations to determine the best performance among groups of learners. Riedesel et al. [40] performed 100 runs of simulations within an application to memorize basic techniques for students. BEETLE II [41] is a simulation-based physics tutor used to foster effective self-explanation in students, requiring 1000 simulation runs to find the best performance using the F-score metric.

In this context, this study contributes to the understanding of the minimum conditions necessary to test the proposed exercise model using simulations and to recreate the possible conditions in a real scenario. In addition, we provide information on the algorithms and minimum exercise interactions needed in the content model design so that other researchers can use these findings in other content such as discussion forums, archives, and wikis, used in any educational system.

III. Base Model for the Characterization of Exercises

Our previous work [13] proposed an exercise model based on interactions performed by students using machine learning algorithms. The model design was named the base model. We selected three characteristics mentioned in related works because the authors used them to characterize an exercise. Although these variables were used earlier, we aimed to understand better the relationship between grade, number of attempts, and time based on previous data on user interactions. The time and grade were based on the student's performance in each attempt. Fig. 1 shows the three characteristics using scatter plots to represent the grade and time for different attempts. The possible values for the three variables are graded with values between 0 and 10, time with values between 0 and n (representing the maximum possible value), and the number of attempts between 1 and m (representing the maximum possible value).

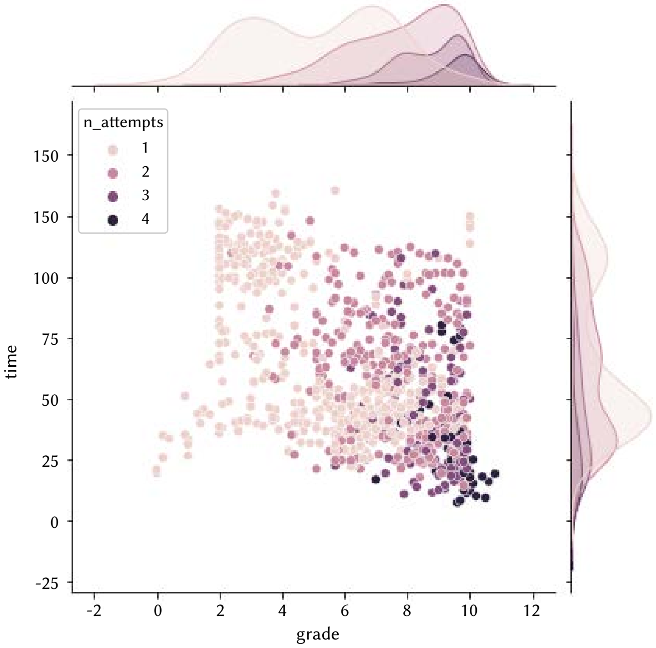
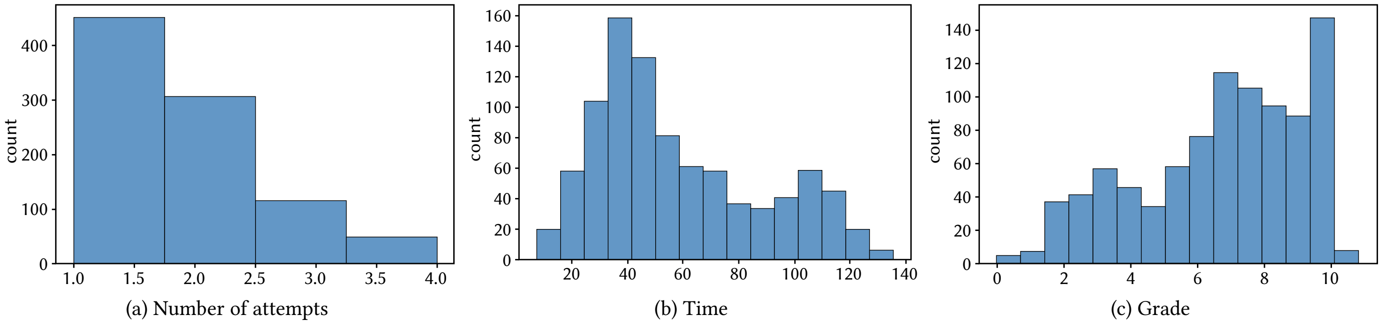


Fig. 1. Characterization of exercises.



(a) Number of attempts (b) Time (c) Grade

Fig. 2. Frequency distribution.

In this study, we used students' simulations to demonstrate their interactions and then trained different intelligence algorithms in the base model and three scenarios by modifying the probability of answering correctly, size of the data set, and distributions.

A. Simulation Using Base Model

We generated a dataset using a normal distribution with specific mean and standard deviation values for each variable including the number of attempts, grade, and time. We consider different criteria when generating each variable and their relationships. For instance, student's grades on the first attempt followed a normal distribution. For subsequent attempts, the same distribution was used but with a limit between the grade of the previous attempt and the maximum possible grade. However, no further attempts were made if students achieved their maximum grades. We used Python programming language with stats and random libraries.

Next, we conducted simulations based on three levels of student knowledge: easy, medium, and difficult. For each level, we adjusted the means and standard deviations of the variables (grade, time, and number of attempts) by increasing or decreasing their values in the distribution function depending on the level of previous knowledge.

We performed at least 150 simulation runs for an exercise with a probability of answering by guessing set at 0% and formed three groups of students categorized as low, medium, and high based on their previous knowledge. Each group comprised 150 students. The mean of the distributions shifted to the left or right depending on the student group. Each student group had a minimum of 150 interactions with the exercise on at least one attempt, and the students were allowed to perform multiple attempts. Simulations were used to train the model and determine the best curve representing the exercise characteristics.

The simulations aimed to recreate possible fictitious cases of student interactions but not to replicate real student behavior. Fig. 2 shows the frequency distribution of the generated variables, which are explained as follows:

- *Number of attempts*: Each student was assumed to have attempted at least one exercise. To preserve the randomness of the data, a random variable was calculated to establish the number of additional attempts that each student performed for that exercise. Subsequently, we performed validations for the second attempt, in which the obtained grade was randomized using a normal distribution, with the minimum value being the grade achieved in the previous attempt. Similarly, for the time variable, we set the randomness using a normal distribution, considering the maximum time obtained in the previous attempt, and ensured that the time did not exceed that of the previous attempt. We followed the same logic for subsequent attempts, such as the third, fourth, and beyond.
- *Grade*: We defined the students' grades obtained during the simulation from 0 to 10. For each attempt, we established a normal distribution, with the mean and standard deviation determined based on the three groups of students during the exercise. All the students had at least one grade for each exercise, as they had attempted it at least once. If students performed multiple attempts at exercise, each grade was obtained using a normal distribution between the maximum possible value for the exercise and the grade obtained on the previous attempt. However, if a student achieved their maximum grade, they were not allowed to make another attempt during the exercise. In all other situations, the new grade depended on students' number of attempts.
- *Time*: The exercise time of the trainees was limited from 0 to m seconds. The specific value of m depends on the difficulty level of the exercise, and in this study, we examined multiple values of m. To ensure that the data remained random, we created a normal distribution with mean and standard deviation values based on the three levels of exercise difficulty. As the number of attempts increases, the time variable decreases. However, as the grade level increased, the time variable also increased. If a student performs multiple exercise attempts, the time obtained is calculated randomly. This value was set as the maximum time calculated in a previous study.

B. Results Using Base Model

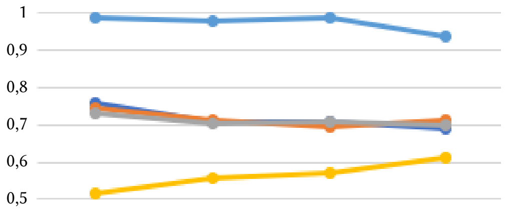
1. Curve Estimation Using Machine Learning

The base model was implemented using machine learning algorithms in Jupyter, using Python v3.9.2 as the programming language. The scikit library, an open-source library that implements many machine learning algorithms, was used. We used the same input dataset for all the algorithms and tested different classifier algorithms using 80% of the data for training and 20% of the simulated data for testing. The classifiers tested included Random Forest (with different depths), Logistic Regression, Nearest Neighbors (with different numbers of neighbors), Gaussian Naive Bayes, and Decision Tree (with different depths). We used grade as the dependent variable and the student's time spent on the exercises and the number of attempts as independent variables. To avoid overfitting, we used cross-validation with an algorithm to obtain the best metrics.

Fig. 3 shows the three best algorithms using precision(macro), recall(macro), f1(macro), RMSE, and AUC as metrics because the data were not balanced. The Nearest Neighbors with the $k = 10$ algorithm obtained relatively good metric values for approximating the relationship between the three variables used in the model design.

C. Best Algorithm Nearest Neighbors

We selected the best algorithm obtained in the previous section(i.e. the Nearest Neighbors with $k = 10$) and the *confusion_matrix* method of the sklearn.metrics library to obtain the confusion matrix. Fig. 4 shows the confusion matrix for the nearest neighbors algorithm with



		Nearst Neighbours (k = 10)	Decision Tree (Max Depth = 10)	Random Forest (Max Depth = 10)	MLP (tanh)
●	Precision (macro)	0,758	0,706	0,710	0,692
●	Recall (macro)	0,746	0,713	0,695	0,714
●	F1 (macro)	0,732	0,702	0,709	0,697
●	RMSE (macro)	0,515	0,558	0,571	0,610
●	AUC	0,99	0,980	0,990	0,940

Fig. 3. Metrics of algorithm model.

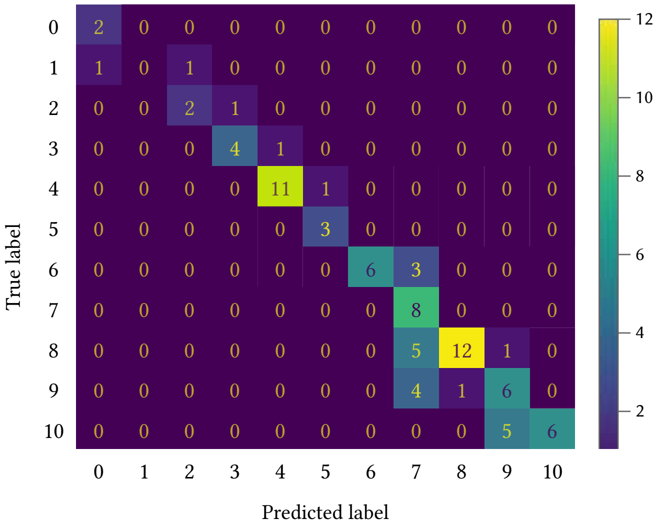


Fig. 4. Confunsion matrix with nearest neighbors(k=10).

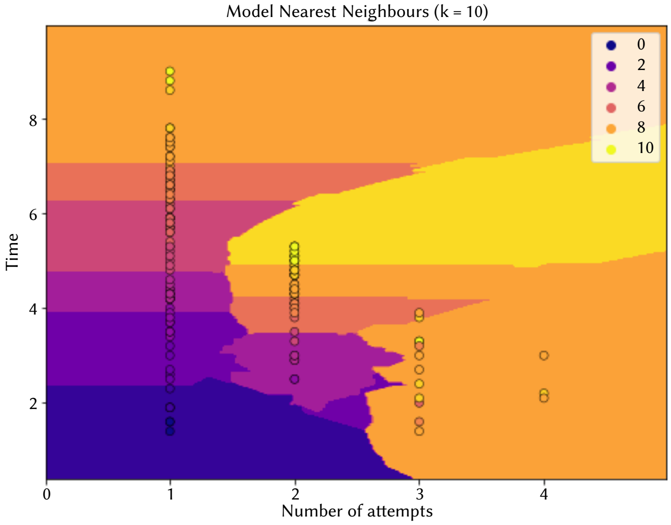


Fig. 5. Plot with nearest neighbors(k=10).

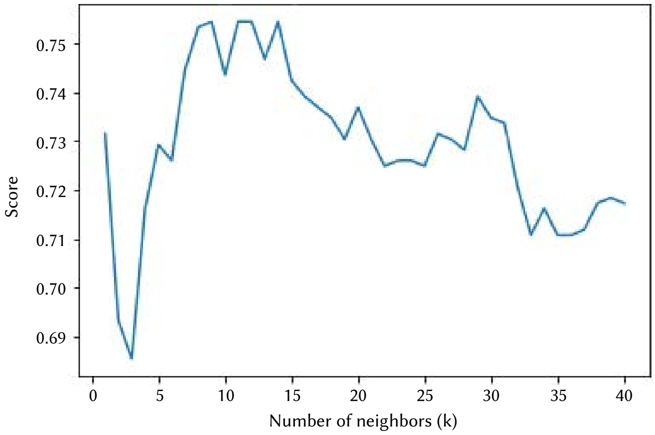


Fig. 6. Cross validation with different numbers of neighbors.

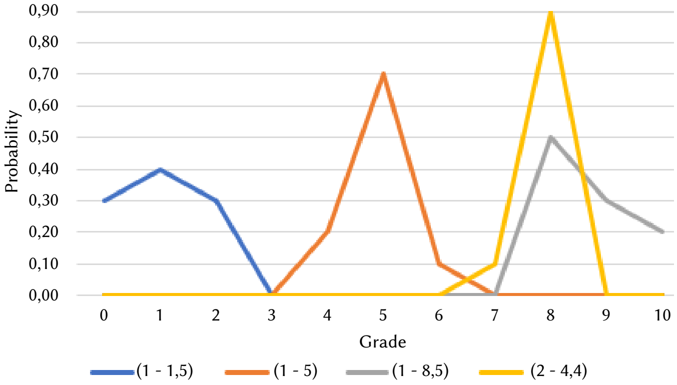


Fig. 7. Predicted probability with nearest neighbors(k=10).

k equal to 10. The matrix shows the different classes for the dependent variable grade of exercise. The figure shows the prediction accuracy for all grades, obtaining the highest values for grades 4 and 6. Also, the lowest accuracy was obtained for grades between 0 and 2.

The different clusters obtained in the model are shown in Fig. 5 for each attempt at different times. The changes between the colors indicate the Bayes decision boundary of the different classes corresponding to the dependent variable grade of colors ranging from blue for 0 to yellow for 10.

We used the *cross_val_score* method from the sklearn.model_selection library to prevent the overfitting of the best algorithm. Fig. 6 shows the results of cross-validation accuracy with five folds evaluated in the nearest neighbor algorithm using different neighbor values. The results indicate that the best values were obtained with neighbor values between 10 and 15, which helped avoid over-fitting and under-fitting. Therefore, the nearest neighbor algorithm with k equal to 10, which obtained the best-evaluated metrics is within this range.

In addition, we used the *predict_proba* method of the KNeighborsClassifier class within the sklearn library to predict the probability of all classes using grade as an independent variable. Fig. 7 shows the results using four elements of the test data set; the X-axis shows the different classes of the variable grade, whereas the Y-axis shows the probability estimate obtained in the predict_proba method. The test input data used were a pair of variables: the first corresponded to the number of attempts, and the second was the time spent.

The results show four examples tested in the selected algorithm in the first scenario (1–1.5): 1 is the number of attempts, and 1.5 is the time spent normalized between 0 and 10, obtaining a cumulative probability of 0.90 for a grade between 0 and 2. Finally, for a second attempt with a time of 4.4, a probability of 0.9 is obtained for the highest grade of 8. In summary, this exercise increased the time spent on the first attempt and the probability of improving grade.

The following sections aim to create different possible scenarios and analyze the performance of the machine learning algorithms to estimate the model using these characteristics. We tested different datasets with a different number of interactions, changing the probability of correctly answering the questions, and changing the distributions, and we used a group of students with the same prior knowledge.

IV. Simulations

In this section, we describe the methodologies used to generate different sets of simulated data for the three simulated datasets for each of the three exercise characteristics.

A. Using Different Probabilities of Answering Correctly

To illustrate the different probabilities of answering a question correctly, we used three types of questions used in student evaluations: 50% to represent true/false questions, 20% and 14% to represent multiple-choice questions with five or seven options and just one correct answer, and 7%and 5% to represent multiple-choice questions with six or seven options and two correct answers. The simulations aimed to understand the model's behavior for different types of questions depending on the probability of answering correctly, identifying changes in the model's accuracy, and whether more interactions are needed.

Using the data simulation, we used the same methods and libraries described in the previous section for the exercise model. We then ran 150 simulations corresponding to the interactions of 150 students during the exercise. Each student completed at least one interaction during the exercise and made more attempts in the same exercise. The grade in each of the simulated exercises was different because it depended on the type of probability.

- 50%: This type of probability corresponds to true/false questions. Students with no prior knowledge had a 50%probability of correctly answering
- 20% and 14%: These two probabilities represent questions with n options, of which only one was the correct answer. We simulated two random questions with a probability of a student answering randomly: 20% (one correct answer out of five options) and 14% (one correct answer out of seven options).
- 7% and 5%: In this type of probability, students had more possible selections because the exercise had a combination of n among m, where n is the number of correct answers and m is the number of choices. For the simulation, we considered two correct answers among the six options; we obtained a combination of 15 possibilities available to the student. Therefore, the probability of correctly answering the questions was 7%(1 out of 15). Finally, the other probability of 5%corresponds to a question with two correct answers among the seven options (1 out of 21).

If the student obtains the maximum grade on the first attempt or N attempts, the student makes no further attempts. The same conditions were used for the simulations in the base model.

B. Using Different Number of Interactions

Initially, we [13] used 150 interactions with a probability of answering by guessing of 0%, and in the previous section, we used 150 interactions with three different probabilities of answering correctly. However, in the present section, we now focus on identifying how large a data set is needed to find the size of the data set needed to find the characteristic curve of the model for this type of question that is accurate enough, similar to what has been done in other studies [42] [43]. To determine the characteristic curve of the model for this type of question, we simulated students' interactions in three exercises with different difficulties based on previous knowledge acquired: low, medium, and high. The data set size options for each exercise were as follows:

- 30 interactions
- 50 interactions
- 100 interactions
- 150 interactions
- 200 interactions
- 300 interactions
- 1000 interactions

C. Using Different Types of Distributions

In a previous work [13], a standard distribution was used for the simulations. However, in the present section, we performed simulations with different distributions for two exercise characteristics: grade, time spent, and the variable number of attempts to keep the distribution fixed in all simulations. The aim was to simulate different student behaviors and identify whether the model fits different possible real-world scenarios. Previous work has used different simulations, [44] used a uniform distribution for the difficulty of questions in simulated student interactions. On the other hand, [45] assumed student ability to be a normal distribution with mean and variance using it to obtain the probability of answering correctly in simulated students. The following distributions were used:

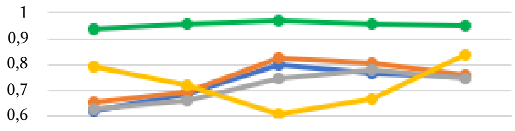
- Uniform distribution: Interactions are centered on intervals (a,b). A possible scenario is that students obtain a similar grade in an exercise, and none have low or high extremes.
- Normal distribution: These are the interactions used in the previous study and previous simulations; it is also the distribution used in other studies [45] [13] where student data were simulated.
- Gamma distribution: Most of the interactions were close to each other, and a few data points were at the end of the bell distribution. For example, almost all students had a similar grade in one exercise, and a few students had a higher grade.

V. Results

A. Using Different Probabilities of Answering Correctly

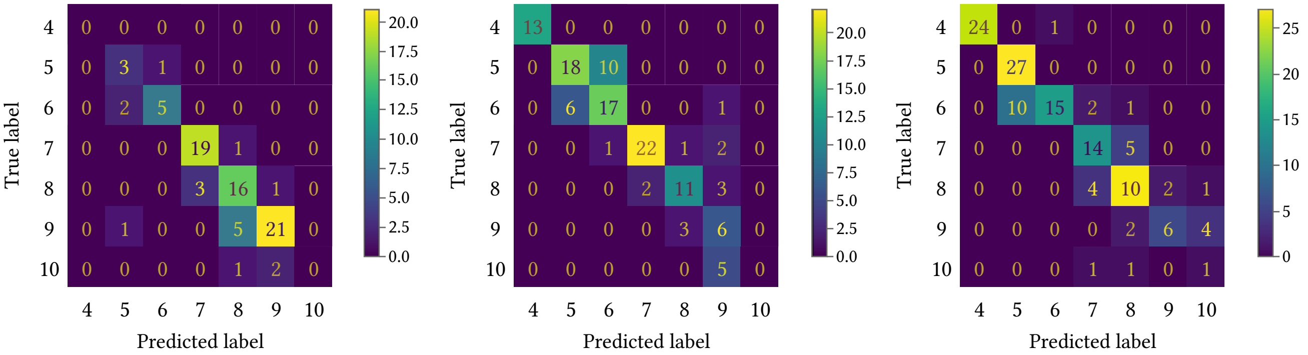
1. Curve Estimation Using Machine Learning

We tested the machine-learning algorithms described in Section III.B. 1 using the metrics previously indicated. As shown in Fig. 8, the best algorithm describing the three different datasets was the nearest neighbor with k equal to 10. The metrics corresponding to the 50% probability have poor results, with values between 0.6 and 0.7, owing to the high probability of answering correctly. Therefore, the model



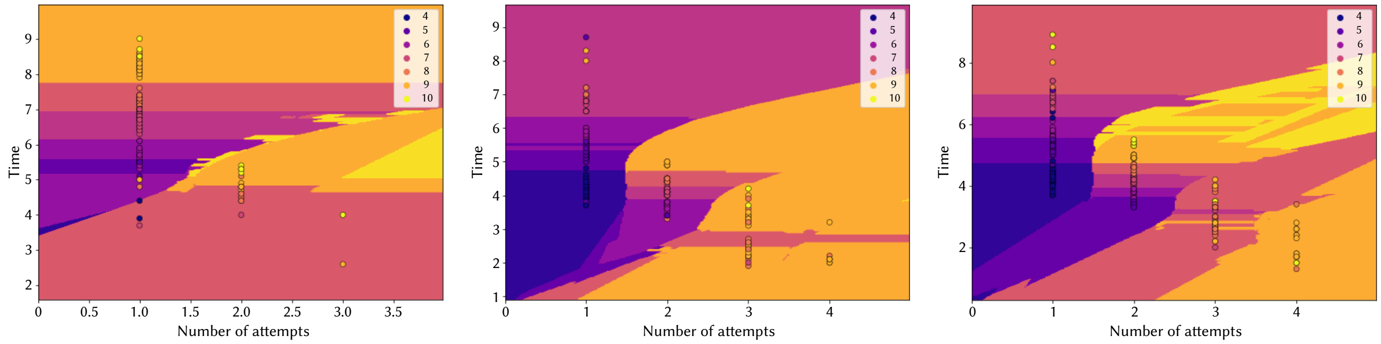
		50% probability	20% probability	14% probability	7% probability	5% probability
●	Precision (macro)	0,619	0,687	0,798	0,769	0,750
●	Recall (macro)	0,652	0,693	0,827	0,806	0,758
●	F1 (macro)	0,628	0,662	0,748	0,777	0,744
●	RMSE (macro)	0,795	0,722	0,610	0,668	0,837
●	AUC	0,940	0,960	0,970	0,960	0,950

Fig. 8. Metric of algorithm Nearest Neighbors (k=10).



(a) (b) (c)

Fig. 9. Confusion matrix with nearest neighbors (k=10).



(a) (b) (c)

Fig. 10. Plot with nearest neighbors(k=10).

had few classes corresponding to lower grades and could not learn correctly for these values. In this scenario, the proposed model design could not be used because it only had two answer choices with a maximum of two attempts to find the correct answer at random by the student. In this type of question with few possible answers, such as True/False, it is not recommended to use the model design because, after two attempts, all students would get the maximum grade, no matter how much time was spent.

In contrast, for probabilities of 5%, 7%, 14%, and 20%, there was a different distribution of grades and sufficient data obtained in each class of the dependent variable for learning the algorithm. By decreasing the probability of answering correctly, the values of all the metrics evaluated improved. The results were unsatisfactory with a probability of 20%, the results were unsatisfactory. In addition, for probabilities of 14%, 7%, and 5%, values greater than 0.75 are obtained in the precision, recall, and F1 metrics, respectively. In contrast, the AUC and RMSE metrics were relatively similar.

In the following subsections, we report the results for three probabilities of answering correctly: 50%, 14%, and 7%, and we select the probability for each type of question.

2. Best Algorithm Nearest Neighbors

Having already identified the best algorithm for predicting the grade, in this subsection, we present three different subsections with three different figures for each probability of answering correctly for the best algorithm. The three scenarios selected for analysis in this study and the following subsections are probabilities of 50%, 14%, and 7%. Fig. 9 shows the three confounding matrices for the algorithm in the three evaluated scenarios. Fig. 9(a) corresponds to a 50% probability of answering; it can be seen that the algorithm has few elements for grades better than 5 and has a test condition for grades 7 and 9. Also, Fig. 9(b) and Fig. 9(c) correspond to 14% and 7% probabilities respectively and Fig. 9(c) has a better accuracy between grades 4 and 7 and presents a particular sensitivity between grades greater than 7. A possible reason may be the small amount of data available for these classes.

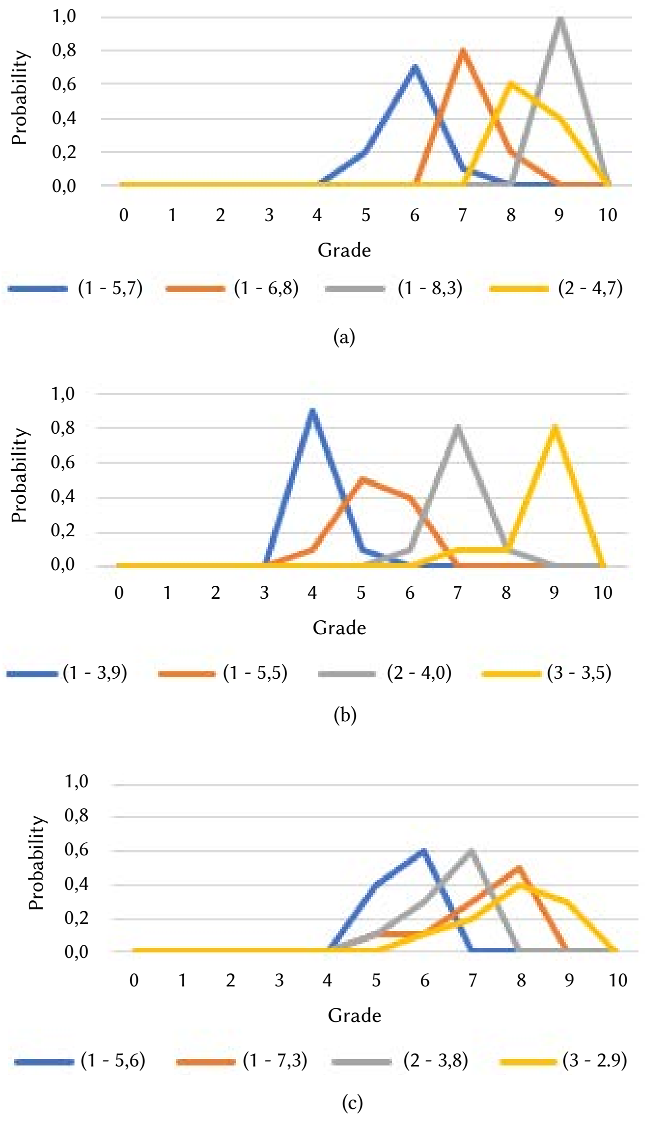
Fig. 10 shows the different clusters corresponding to the nearest neighbor algorithm with k equal to 10. The limits of each class vary depending on the percentage of probability of answering correctly. The tonality varies with color to yellow, corresponding to class 10 in the question with 50% (Fig. 10(a)), while 14% (Fig. 10(b)) and 7% (Fig. 10(c)) show the whole range of tonality from the blue of class 0 to the yellow color corresponding to class 10.

To avoid overfitting, we evaluated the score of the algorithm using the *cross_val_score* of the sklearn.model_selection library with accuracy as scoring and 5-fold cross-validation. Table I shows the results of the cross-validation performed with five subsets of the nearest neighbors algorithm with the three different datasets representing the three types of questions. The N-fold column indicates the run number and the value indicates the accuracy of the algorithm in this run. Thus, we evaluated the robustness of the algorithm and avoided overfitting.

TABLE I. Cross-Validation Values

	50% probability	**14% probability**	**7% probability**
cv1	0,679	0,799	0,768
cv2	0,575	0,826	0,793
cv3	0,616	0,812	0,692
cv4	0,676	0,740	0,781
cv5	0,548	0,809	0,806

Finally, we used the *predict_proba* method to calculate the probability of different grades using the simulated test dataset with three different probabilities of answering correctly. For example, in Fig. 11, using in the model a similar ordered pair, such as (1–5.5), (1–5.6), or (1–5.7), where 1 means the number of attempts and 5.5,5.6,5.7



(a)

(b)

(c)

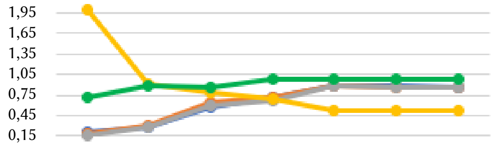
Fig. 11. Predicted probability with nearest neighbors(k=10).

corresponds to the time spent in the exercise. We obtained a probability of 0.7 for a grade of 6 in the exercise with 50%correct answers (Fig. 11(a)). In comparison, with an exercise of 14% (Fig. 11(b)), we obtained a probability of 0.9 for a grade of 4, and an exercise of 7% (Fig. 11(c)), we obtained a 0.6 probability for a grade of 6. As we can observe, we obtained different probabilities in the classes for the three probabilities of answering correctly, evaluated with similar values of time spent in the first attempt.

B. Using Different Number of Interactions

1. Curve Estimation Using Machine Learning

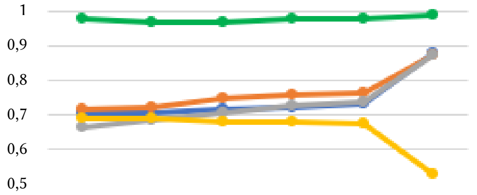
We tested the same algorithm used in the previous section, using the same metrics. Fig. 12 shows the precision, recall, F1, RMSE, and AUC metrics. We can see an increase in their values as the number of interactions increased, stabilizing the curve at 200 interactions. The RMSE metric decreased as the number of interactions increased, achieving stability with the same number of interactions as that of the other metrics. From the results, we can conclude that the minimum number of interactions for the proposed exercise model with good accuracy is approximately 200 because the best results were obtained for all the metrics evaluated: precision of 0.878, recall of 0.873, F1 of 0.875, and RMSE of 0.527.



		30 interactions	50 interactions	100 interactions	150 interactions	200 interactions	300 interactions	1000 interactions
●	Precision (macro)	0,214	0,269	0,568	0,708	0,878	0,882	0,866
●	Recall (macro)	0,19	0,315	0,632	0,718	0,873	0,862	0,871
●	F1 (macro)	0,167	0,281	0,587	0,666	0,875	0,865	0,864
●	RMSE (macro)	2,000	0,918	0,793	0,693	0,527	0,526	0,527
●	AUC	0,723	0,890	0,870	0,980	0,990	0,985	0,989

Fig. 12. Metric of algorithm Nearest Neighbors (k=10).

Regarding the previous result, we performed simulations with values between 150 and 200 interactions to determine the exact value at which the curve of the metrics stabilizes. Fig. 13 shows the results of all the metrics evaluated, and the results show that between the values of 150 and 190 interactions, the values of the metrics precision, recall, F1, and RMSE increase slightly. For the 200 interactions, all metric values yielded the best results, as indicated in the previous paragraph. In summary, for the simulated exercise with values for the three characteristics, the number of attempts (mean, 2; minimum, 1; maximum, 4) of 200 interactions was needed.

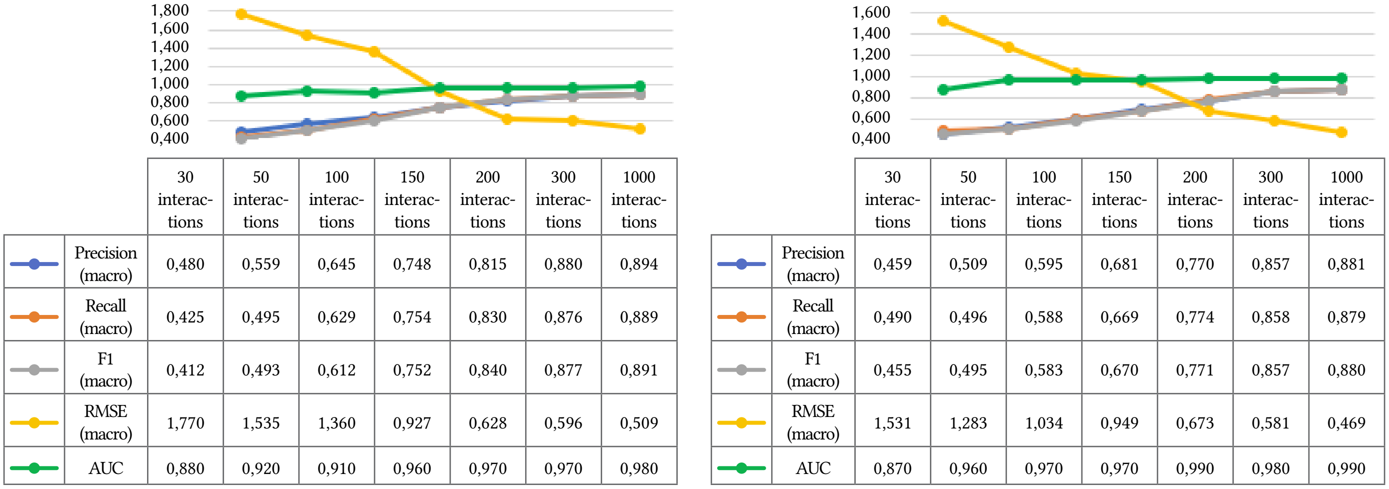


		150 interactions	160 interactions	170 interactions	180 interactions	190 interactions	200 interactions
●	Precision (macro)	0,708	0,709	0,716	0,721	0,731	0,878
●	Recall (macro)	0,718	0,725	0,747	0,757	0,764	0,873
●	F1 (macro)	0,666	0,687	0,706	0,727	0,737	0,875
●	RMSE (macro)	0,693	0,692	0,683	0,683	0,674	0,527
●	AUC	0,98	0,970	0,970	0,980	0,980	0,990

Fig. 13. Metric of algorithm Nearest Neighbors (k=10) between 150 and 200.

If the number of attempts to obtain the maximum grade increases, then. To test this hypothesis, we simulated two new exercises by modifying the variable number of attempts differently from the previous exercise with a mean of two. The first exercise involved an average of four attempts, and the second exercise involved an average of seven attempts.

Fig. 14(a) shows the results of the precision, recall, F1, RMSE, and AUC metrics for different interactions in the exercise with a mean of four attempts. The metrics decreased as the number of interactions increased, except for the AUC metric, which tended to maintain similar values. Fig. 14 shows that in 500 interactions, good values were obtained for all metrics evaluated: *precision* = 0.815, *recall* = 0.83, *F1* = 0.84, *RMSE* = 0.628, and *AUC* = 0.92. We conclude that, for an exercise with a mean of four attempts to find the maximum grade, a minimum of 500 interactions are needed.



		30 interactions	50 interactions	100 interactions	150 interactions	200 interactions	300 interactions	1000 interactions
🟦	Precision (macro)	0,480	0,559	0,645	0,748	0,815	0,880	0,894
🟧	Recall (macro)	0,425	0,495	0,629	0,754	0,830	0,876	0,889
⬜	F1 (macro)	0,412	0,493	0,612	0,752	0,840	0,877	0,891
🟨	RMSE (macro)	1,770	1,535	1,360	0,927	0,628	0,596	0,509
🟩	AUC	0,880	0,920	0,910	0,960	0,970	0,970	0,980

(a)

		30 interactions	50 interactions	100 interactions	150 interactions	200 interactions	300 interactions	1000 interactions
🟦	Precision (macro)	0,459	0,509	0,595	0,681	0,770	0,857	0,881
🟧	Recall (macro)	0,490	0,496	0,588	0,669	0,774	0,858	0,879
⬜	F1 (macro)	0,455	0,495	0,583	0,670	0,771	0,857	0,880
🟨	RMSE (macro)	1,531	1,283	1,034	0,949	0,673	0,581	0,469
🟩	AUC	0,870	0,960	0,970	0,970	0,990	0,980	0,990

(b)

Fig. 14. Metric of algorithm Nearest Neighbors (k=10) increasing the number of attempts.

Furthermore, Fig. 14 (b) shows the results of the same metrics in the exercise with a mean of seven attempts. In the exercise with a mean of four attempts, the metrics decreased as the number of interactions increased, except for the AUC metric. Fig. 14(b) shows that at 800 interactions, good values were obtained for all metrics: *precision* = 0.875, *recall* = 0.858, *F1* = 0.857, *RMSE* = 0.581, and *AUC* = 0.980. Based on these results, we can conclude that, in an exercise with a mean of seven attempts, a minimum of 800 interactions would be needed. In summary, as the number of attempts that students must make to obtain the maximum grade increases, the minimum number of simulated students also increases.

2. Best Algorithm Nearest Neighbors (K=10) With 200 Interactions

Fig. 15 shows the confusion matrix of the algorithm with the highest accuracy with 200 interactions obtained in the previous subsection. The results show that the algorithm has a better prediction for middle grades, with a decreasing prediction as the grades increase. However, for low grades, the accuracy decreases because of the small dataset with which the model was trained and because of the type of distribution used in the simulations.

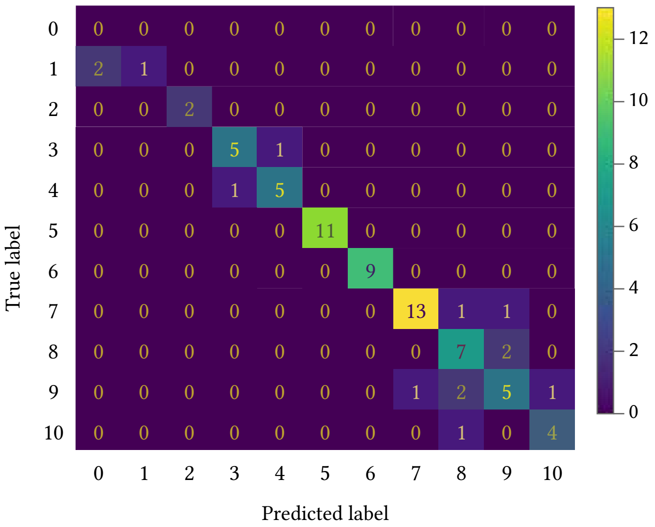


Fig. 15. Confusion matrix with nearest neighbors(k=10).

Additionally, Fig. 16 shows different clusters of the algorithm for different numbers of attempts. The tonalities of the different classes varied as number of attempts increased, the students improved their grades, and the total number of classes decreased. In the middle grade, most of the clusters were located in the correct class on the first attempt. In the low and high grades, the classes had clusters corresponding to the nearby classes.

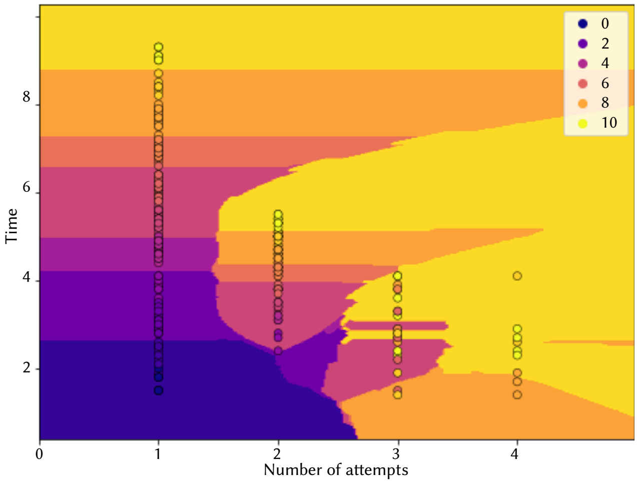


Fig. 16. Plot with nearest neighbors(k=10).

Using the same method as in Section III.A.2, Table II lists the variables considered in the model with their possible values. The N-fold column indicates the run number and the value indicates the accuracy of the algorithm in this run. The results show good accuracy of the algorithm in the five different subsets of the cross-validation, indicating the robustness of the algorithm and the avoidance of overfitting.

TABLE II. Cross-Validation Values

N-fold	value
cv1	0,892
cv2	0,965
cv3	0,862
cv4	0,789
cv5	0,862

Finally, we used the *predict_proba* method to calculate the probability of different grades using the simulated test data. Fig. 17 shows the results obtained using these four examples. The first example (1-3.0) corresponds to the first attempt in a time of 3 units, which corresponds to a higher probability of obtaining a grade of 2. In contrast, (3–3.1) corresponds to the third attempt at a time of 3.1. Similar to the previous study, the results vary, obtaining a higher probability for a grade of 7 or 10. In conclusion, according to the predictions of this exercise, if a student spends more time during the first attempt or makes more than one attempt, the student has a greater probability of obtaining a higher grade.

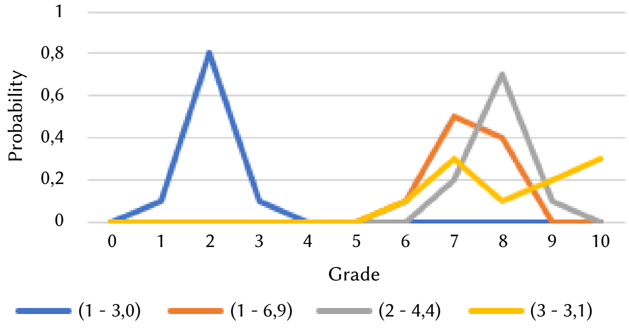
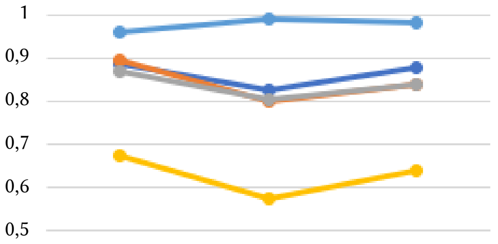


Fig. 17. Predicted probability with nearest neighbors(k=10).

C. Using Different Types of Distributions

1. Curve Estimation Using Machine Learning

To answer RQ4, we trained the model using the three datasets corresponding to each distribution. The results were obtained using the same algorithm as that in the previous section, and the same metrics are shown in Fig. 18. In general, for the three different student behaviors (three distributions), the precision, recall, and F1 metrics exhibited values ranging from 0,8 and 0,9. In contrast, the RMSE decreased slightly when the student's behavior was normally distributed. Finally, the AUC of the three distributions was not significantly different, with a value very close to 1. In conclusion, the different distributions of student behavior using machine learning algorithms converged with good results.

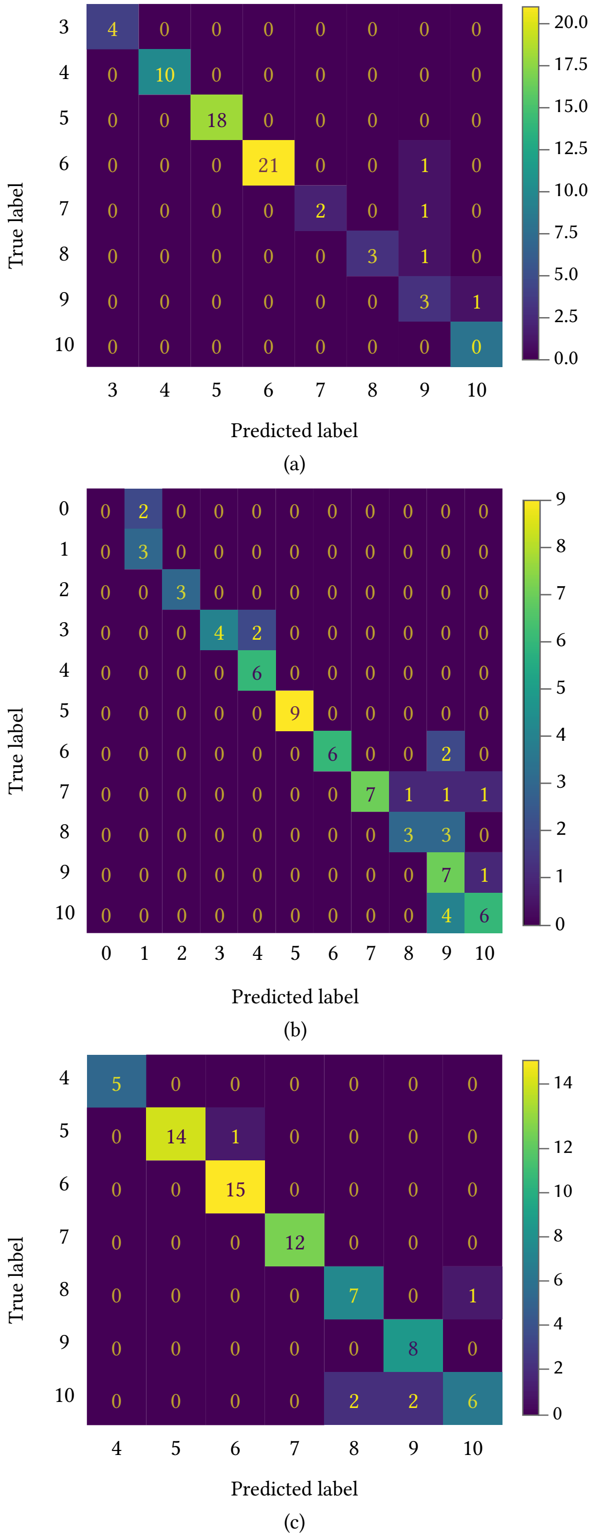


		Uniform distribution	Normal distribution	Gamma distribution
●	Precision (macro)	0,885	0,827	0,875
●	Recall (macro)	0,896	0,801	0,840
●	F1 (macro)	0,869	0,804	0,836
●	RMSE (macro)	0,672	0,574	0,639
●	AUC	0,96	0,990	0,980

Fig. 18. Metric of algorithm Nearest Neighbors (k=10).

2. Best Algorithm Nearest Neighbors (K=10)

The confusion matrix allowed us to observe the behavior of the algorithm by relating the preconditions to the real cases. Fig. 19 shows the confusion matrices for the three different distributions used. Fig. 19(a) corresponds to a uniform distribution; Fig. 19(b) corresponds to a normal distribution; and Fig. 19(c) corresponds to a gamma distribution. Fig. 19(a) and Fig. 19(c) show that not all values are available for the degree of the dependent variable. By contrast, in Fig. 19(b), all grade classes can be obtained. Moreover, Fig. 19(a) and Fig. 19(c) show similar behavior, obtaining a high precision for the mean grades, whereas Fig. 19(b) shows good precision distributed over a larger number of grades.



	3	4	5	6	7	8	9	10
3	4	0	0	0	0	0	0	0
4	0	10	0	0	0	0	0	0
5	0	0	18	0	0	0	0	0
6	0	0	0	21	0	0	1	0
7	0	0	0	0	2	0	1	0
8	0	0	0	0	0	3	1	0
9	0	0	0	0	0	0	3	1
10	0	0	0	0	0	0	0	0

(a)

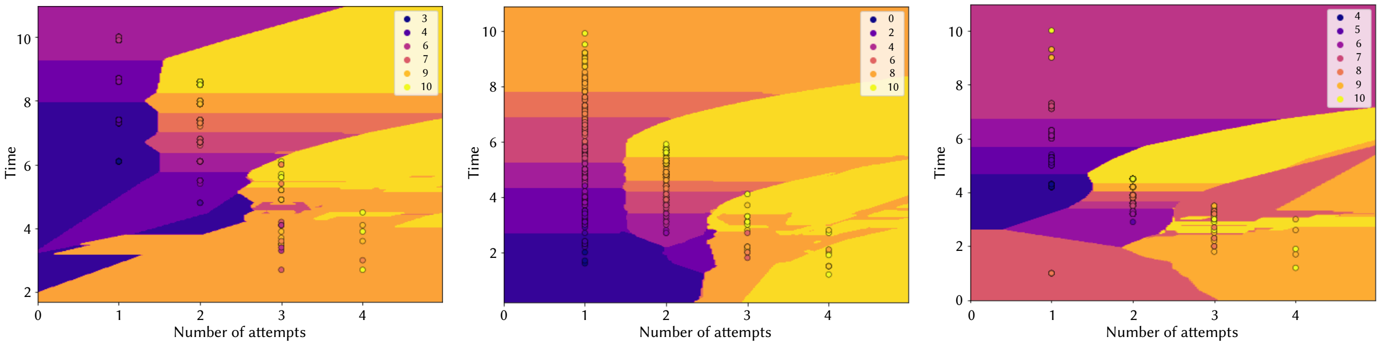
	0	1	2	3	4	5	6	7	8	9	10
0	0	2	0	0	0	0	0	0	0	0	0
1	0	3	0	0	0	0	0	0	0	0	0
2	0	0	3	0	0	0	0	0	0	0	0
3	0	0	0	4	2	0	0	0	0	0	0
4	0	0	0	0	6	0	0	0	0	0	0
5	0	0	0	0	0	9	0	0	0	0	0
6	0	0	0	0	0	0	6	0	0	2	0
7	0	0	0	0	0	0	0	7	1	1	1
8	0	0	0	0	0	0	0	0	3	3	0
9	0	0	0	0	0	0	0	0	0	7	1
10	0	0	0	0	0	0	0	0	0	4	6

(b)

	4	5	6	7	8	9	10
4	5	0	0	0	0	0	0
5	0	14	1	0	0	0	0
6	0	0	15	0	0	0	0
7	0	0	0	12	0	0	0
8	0	0	0	0	7	0	1
9	0	0	0	0	0	8	0
10	0	0	0	0	2	2	6

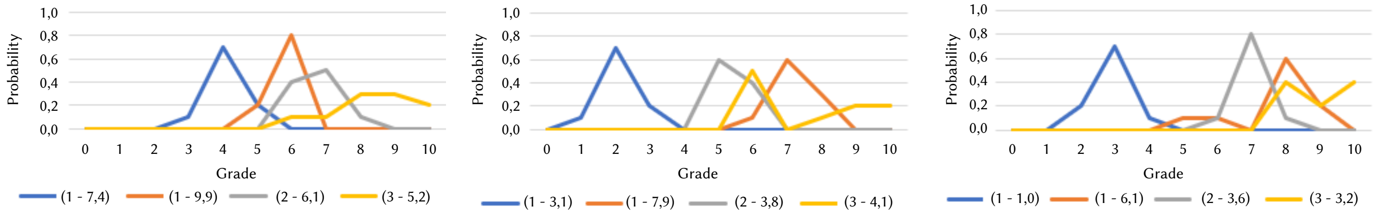
(c)

Fig. 19. Confunsion matrix with nearest neighbors(k=10).



(a) (b) (c)

Fig. 20. Plot with nearest neighbors(k=10).



(a) (b) (c)

Fig. 21. Predicted probability with nearest neighbors(k=10).

Next, Fig. 20 shows the different clusters of the dependent variable using the nearest neighbor algorithm with k equal to 10. Fig. 20(a) corresponds to the uniform distribution, Fig. 20(b) a normal distribution and Fig. 20 for the gamma distribution. The three figures represent three different student behaviors, where the different shades represent each class of the grade variable. The dispersion of the clusters in the different attempts was related to the type of distribution used, as shown in Fig. 20(a) and Fig. 20(c), with a large dispersion in each attempt. In addition, Fig. 20(b) shows a better distribution of clusters in each class, as represented by the same colors.

Using the same method as in Section III.A.2, Table III presents the results of the cross-validation performed with three subsets representing the three types of distributions. The N-fold column indicates the run number and the value indicates the accuracy of the algorithm in this run. Accuracy of different subsets in cross-validation obtained good results, demonstrating the robustness of the algorithm in different distributions and avoiding overfitting of the algorithm.

TABLE III. Cross-Validation Values

N-fold	Uniform distribution	Normal distribution	Gamma distribution
cv1	0,849	0,824	0,887
cv2	0,876	0,810	0,928
cv3	0,917	0,838	0,914
cv4	0,903	0,796	0,818
cv5	0,876	0,867	0,832

Finally, we used the *predict_prob method* to calculate the probability of different grades, and the dataset used as a test was a part of the simulated data. Fig. 21 shows the results obtained using three different simulations. Fig. 21(a) corresponds to a uniform distribution, Fig. 21(b) corresponds to a normal distribution, and Fig. 21(c) corresponds to a gamma distribution. In conclusion, the probabilities obtained had different values for the three evaluated datasets.

VI. Discussion

In this section, we analyze the results obtained from the simulations based on our research questions.

A. RQ1: Is It Possible to Obtain a Time-Grade-Attempts Model in Exercises That Are Accurate Enough Using Some Traditional Machine Learning Algorithms?

The results show that traditional machine learning algorithms can model exercises using independent variables, such as time and number of attempts, with the dependent variable being the grade obtained by the student. The four algorithms that could be used were nearest neighbor (k=10), Decision Tree (Max Depth=10), Random Forest (Max Depth=10), and MLP (tanh). We evaluated the effectiveness of all algorithms based on metrics such as precision, recall, F1, RMSE, and AUC and found that these four algorithms yielded the best results. Finally, we recommend using the nearest neighbor algorithm (k=10) because the first choice, as it achieved the best results in the simulations conducted. This algorithm and its variants have been used in various applications such as medical predictions, data mining, and financial modeling [46].

B. RQ2: How Does the Accuracy of the Models Vary for Different Types of Exercises?

The findings show a variation in the values of the metrics evaluated for the different questions. Modifying the probability of answering correctly implies obtaining different data dispersions among dependent variable classes (grades). First, the questions with a 50% probability had the worst result among the others. This is because of the high probability of obtaining a good grade randomly even if the student has no prior knowledge. As the probability of answering correctly decreases, better results are obtained in the metrics because of the data distribution, and the algorithm has the necessary information to learn correctly. We do not recommend using the proposed model for true/false questions corresponding to 50% probability because of its low effectiveness and the limited data that will be obtained regarding the number of student attempts.

By contrast, the model has better results in the metrics using an artificial intelligence algorithm for the type of question with multiple options represented by probabilities of 14%, 7%, and 5%. For 20% probability, the results show a small increase in the results for 50% probability. There is an inverse relationship between probability and metrics; as the probability of answering correctly increases, the values of the evaluated metrics decrease.

This type of question is perceived as better and preferred by the students [47]. Using the information obtained from the model, teachers can orchestrate the process by redesigning educational exercises to improve student learning, as in other studies [48] [49]. For example, by knowing the types of questions, teachers can modify their exams based on student's grades, the number of attempts that students will have to make, and the time it will take to finish the questions. By using this information, teachers can redesign questions with better results based on the proposed model to improve students' learning processes.

C. RQ3: What Minimum Number of Interactions Is Required to Stabilize the Exercise Model With Acceptable Accuracy?

Previous studies [42] [43] examined the impact of various sample sizes on model stability and accuracy, to identify the minimum number of sizes required to optimize the characteristic curve. The results show that we need a minimum of 200 student interactions in the exercise to model the three characteristics of the proposed model design for exercises, with an average of two attempts to obtain the maximum grade.

However, we could also use 300 or more interactions for the first attempt because the difference in accuracy was insignificant because there were few classes to classify. The accuracy of the algorithm did not increase significantly in the model considering the threshold of 200 interactions as the number of interactions increased. Having a minimum of 200 interactions performed in an exercise does not necessarily imply having 200 students because the same student can perform multiple interactions when trying to solve the same exercise several times, which increases the total number of interactions.

Moreover, if the exercises require more attempts to obtain the maximum grade, such as 4 or 7, more interactions are required to converge the model. The findings showed that we would need a minimum of 500 and 800 interactions for these two types of exercises. Existing a direct relation between the number of attempts and the number of interjections, if the number of attempts needed to obtain the maximum score increases, the number of interactions will be higher.

The results can be used to analyze any platform on which the proposed exercise model should be tested: for example, in a massive open online course (MOOC), because of the large number of learners and the possibility of obtaining numerous interactions; or in contrast, in Learning Management System (LMS) courses with a specific number of learners.

D. RQ4: How Do Different Forms of Student Behavior Modify the Results Obtained in the Previous Research Questions?

We used different distributions in studies of students with exercise characteristics. For example, normal distributions have been used to address question difficulty and student skills [50]. Once we identified the number of interactions and type of questions required to estimate the model and obtained good results for all metrics, we ran simulations with different distributions to recreate possible student scenarios. For example, students obtained an average grade on the first attempt; however, after several attempts, they could not improve their grades. No matter how many attempts the students made, they could not get the maximum grade, or all students achieved high grades on their first attempt. The findings allow us to infer that we can adapt the proposed model to different scenarios to help teachers identify the problems that students face when solving the exercise and redesign the exercise if necessary to improve the grade obtained by the student.

VII. Conclusion and Future Work

The simulations allowed us to illustrate the exercise model under different scenarios. First, we found that a traditional machine learning algorithm could model the exercise while obtaining acceptable results for the metrics evaluated, and the robustness of the model was evaluated using five-field cross-validation. Next, we found that different probabilities of answering a question correctly affect the accuracy of the model due to the distribution of the scores obtained as a function of the probability; for questions with few answer options, we do not recommend using the proposed model design as in the case of a true/false question.

Moreover, identifying the number of interactions is essential for testing the model because it indicates the minimum number of students required to evaluate the model accurately. The findings indicated that the number of interactions is related to the number of attempts required to obtain the maximum grade. For example, model design requires a minimum of 200 interactions for an exercise, with an average of two attempts. However, if the number of attempts is increased, more interactions will be required to converge the model.

Also, the model design converges on the different interaction scenarios of the students simulating different student behaviors using three different distributions so that future work can evaluate the model with other distributions.

The present study's findings will allow teachers to redesign their course exercises, knowing information about the three characteristics of exercise: number of attempts, time, and grade for each type of question. For example, identifying student patterns in exercises, such as students failing to improve their grades after a few attempts or investing too much time to improve their grades, among others.

A limitation of the present study is the use of simulated data rather than real data, which does not allow us to generalize the accuracy obtained by applying the different algorithms tested for the three exercise characteristics. However, the results obtained allowed us to illustrate the behavior of the model in the different scenarios evaluated and identify the features needed in the test design in a real environment.

In future work, we plan to test the model in a real-world scenario based on simulation results. This would require many students to interact with the educational exercise. Hence, a massive course is necessary to evaluate the generality of the proposed model. Additionally, we will create explainable visualizations with information about the model within the dataset, that the teacher can use to detect exercise patterns using the three indicated characteristics. Moreover, the teacher can redesign the exercise based on these visualizations to enhance the students' learning.

Acknowledgment

This work was supported in part by the FEDER/Ministerio de Ciencia, Innovación y Universidades–Agencia Estatal de Investigación, through the Smartlet Project under Grant TIN2017-85179-C3-1-R and the H2O Learn Project under Grant PID2020-112584 RB-C31, in part by the Madrid Regional Government through the e-Madrid-CM Project under Grant S2018/TCS-4307 a project which is co-funded by the European Structural Funds (FSE and FEDER).

References

[1] P. Brusilovsky, S. Edwards, A. Kumar, L. Malmi, L. Benotti, D. Buck, P. Ihantola, R. Prince, T. Sirkiä, S. Sosnovsky, *et al.*, "Increasing adoption of smart learning content for computer science education," in *Proceedings of the Working Group Reports of the 2014 on Innovation & Technology in Computer Science Education Conference*, 2014, pp. 31–57.

[2] E. G. Rincon-Flores, E. Lopez-Camacho, J. Mena, O. Olmos, "Teaching through learning analytics: Predicting student learning profiles in a physics course at a higher education institution," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 7, pp. 82–89, 2022.

[3] M. O. Edelen, B. B. Reeve, "Applying item response theory (irt) modeling to questionnaire development, evaluation, and refinement," *Quality of Life Research*, vol. 16, no. 1, pp. 5–18, 2007.

[4] I. Rushkin, I. Chuang, D. Tingley, "Modelling and using response times in online courses," *arXiv preprint arXiv:1801.07618*, 2018.

[5] A. Jiménez-Macías, P. J. Muñoz-Merino, M. Ortiz- Rojas, M. Muñoz-Organero, C. Delgado Kloos, "Content modeling in smart learning environments: A systematic literature review," *Journal of Universal Computer Science (JUCS)*, vol. 30, no. 3, pp. 333–362, 2024.

[6] P. M. Moreno-Marcos, D. M. de la Torre, G. G. Castro, P. J. Muñoz-Merino, C. D. Kloos, "Should we consider efficiency and constancy for adaptation in intelligent tutoring systems?," in *International Conference on Intelligent Tutoring Systems*, 2020, pp. 237–247, Springer.

[7] M. Feng, J. Beck, N. Heffernan, K. Koedinger, "Can an intelligent tutoring system predict math proficiency as well as a standardized test?," in *Proceedings of the 1st International Conference on Education Data Mining*, 2008, pp. 107–116.

[8] B. Martin, A. Mitrovic, K. R. Koedinger, S. Mathan, "Evaluating and improving adaptive educational systems with learning curves," *User Modeling and User-Adapted Interaction*, vol. 21, pp. 249–283, 2011.

[9] F. Dorça, "Implementation and use of simulated students for test and validation of new adaptive educational systems: A practical insight," *International Journal of Artificial Intelligence in Education*, vol. 25, no. 3, pp. 319–345, 2015.

[10] E. Poitras, Z. Mayne, L. Huang, T. Doleck, L. Udy, S. Lajoie, "Simulated student behaviors with intelligent tutoring systems: Applications for authoring and evaluating network-based tutors," *Tutoring and Intelligent Tutoring Systems. Nova Publishers*, 2018.

[11] J. Champaign, R. Cohen, "A model for content sequencing in intelligent tutoring systems based on the ecological approach and its validation through simulated students," in *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010)*, 2010, pp. 486–491.

[12] R. Pelánek, "Metrics for evaluation of student models.," *Journal of Educational Data Mining*, vol. 7, no. 2, pp. 1–19, 2015.

[13] A. Jiménez-Macías, P. J. Muñoz-Merino, C. Delgado Kloos, "A model to characterize exercises using probabilistic methods," in *Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'21)*, 2021, pp. 594–599.

[14] J. M. Spector, "Smart learning environments: Concepts and issues," Society for Information Technology & teacher education international conference, 2016, pp. 2728–2737, Association for the Advancement of Computing in Education (AACE).

[15] E. Pecheanu, C. Segal, D. Stefanescu, "Content modeling in intelligent instructional environments," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, 2003, pp. 1229–1234, Springer.

[16] J. P. Lalor, H. Wu, H. Yu, "Learning latent parameters without human response patterns: Item response theory with artificial crowds," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2019, 2019, p. 4240, NIH Public Access.

[17] F. Martínez-Plumed, R. B. Prudêncio, A. Martínez- Usó, J. Hernández-Orallo, "Item response theory in ai: Analysing machine learning classifiers at the instance level," *Artificial Intelligence*, vol. 271, pp. 18–42, 2019.

[18] C.-M. Chen, H.-M. Lee, Y.-H. Chen, "Personalized e-learning system using item response theory," *Computers & Education*, vol. 44, no. 3, pp. 237–255, 2005.

[19] D. Abbakumov, "The solution of the "cold start problem" in e-learning," *Procedia-Social and Behavioral Sciences*, vol. 112, pp. 1225–1231, 2014.

[20] K. Xue, V. Yaneva, C. Runyon, P. Baldwin, "Predicting the difficulty and response time of multiple choice questions using transfer learning," in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2020, pp. 193–197.

[21] V. Yaneva, P. Baldwin, J. Mee, *et al.*, "Predicting the difficulty of multiple choice questions in a high- stakes medical exam," Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, 2019, pp. 11–20.

[22] Z. Qiu, X. Wu, W. Fan, "Question difficulty prediction for multiple choice problems in medical exams," Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 139–148.

[23] L. Benedetto, A. Cappelli, R. Turrin, P. Cremonesi, "R2de: a nlp approach to estimating irt parameters of newly generated questions," in *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 2020, pp. 412–421.

[24] B. A. Lehman, D. Zapata-Rivera, "Student emotions in conversation-based assessments," *IEEE Transactions on Learning Technologies*, vol. 11, no. 1, pp. 41–53, 2018.

[25] N. Capuano, S. Caballé, J. Conesa, A. Greco, "Attention-based hierarchical recurrent neural networks for mooc forum posts analysis," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–13, 2020.

[26] T. Atapattu, K. Falkner, M. Thilakaratne, L. Sivaneasharajah, R. Jayashanka, "What do linguistic expressions tell us about learners' confusion? a domain-independent analysis in moocs," *IEEE Transactions on Learning Technologies*, vol. 13, no. 4, pp. 878–888, 2020.

[27] M. Feng, N. Heffernan, K. Koedinger, "Addressing the assessment challenge with an online system that tutors as it assesses," *User modeling and user-adapted interaction*, vol. 19, no. 3, pp. 243–266, 2009.

[28] E. Verdú, M. J. Verdú, L. M. Regueras, J. P. de Castro, R. García, "A genetic fuzzy expert system for automatic question classification in a competitive learning environment," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7471–7478, 2012.

[29] M. Uto, "Rater-effect irt model integrating supervised lda for accurate measurement of essay writing ability," in *International Conference on Artificial Intelligence in Education*, 2019, pp. 494–506, Springer.

[30] H. A.-M. Gerlache, P. M. Ger, L. de la Fuente Valentín, "Towards the grade's prediction. a study of different machine learning approaches to predict grades from student interaction data," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, pp. 196–204, 2022.

[31] K. VanLehn, S. Ohlsson, R. Nason, "Applications of simulated students: An exploration," *Journal of artificial intelligence in education*, vol. 5, pp. 135–135, 1994.

[32] N. Matsuda, W. W. Cohen, K. R. Koedinger, "Teaching the teacher: tutoring simstudent leads to more effective cognitive tutor authoring," *International Journal of Artificial Intelligence in Education*, vol. 25, no. 1, pp. 1– 34, 2015.

[33] S. B. Blessing, "A programming by demonstration authoring tool for model-tracing tutors," *International Journal of Artificial Intelligence in Education*, vol. 8, no. 3- 4, pp. 233–261, 1997.

[34] D. E. K. Lelei, G. McCalla, "Simulation in support of lifelong learning design: A prospectus.," in *SLLL@ AIED*, 2019, pp. 38–42.

[35] A. C. Graesser, "Conversations with autotutor help students learn," *International Journal of Artificial Intelligence in Education*, vol. 26, no. 1, pp. 124–132, 2016.

[36] A. Vizcaíno, "A simulated student can improve collaborative learning," *International Journal of Artificial Intelligence in Education*, vol. 15, no. 1, pp. 3–40, 2005.

[37] D. E. K. Lelei, G. McCalla, "How many times should a pedagogical agent simulation model be run?," in *International Conference on Artificial Intelligence in Education*, 2019, pp. 182–193, Springer.

[38] G. Erickson, S. Frost, S. Bateman, G. McCalla, "Using the ecological approach to create simulations of learning environments," in *Artificial Intelligence in Education*, 2013, pp. 411–420, Springer.

[39] S. Frost, G. McCalla, "Exploring through simulation an instructional planner for dynamic open-ended learning environments," in *Artificial Intelligence in Education*, 2015, pp. 578–581, Springer.

[40] M. A. Riedesel, N. Zimmerman, R. Baker, T. Titchener, J. Cooper, "Using a model for learning and memory to simulate learner response in spaced practice," in *Artificial Intelligence in Education*, 2017, pp. 644–649,

Springer.

[41] M. Dzikovska, N. Steinhauser, E. Farrow, J. Moore, G. Campbell, "Beetle ii: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics," *International Journal of Artificial Intelligence in Education*, vol. 24, pp. 284–334, 2014.

[42] T. R. O'Neill, J. L. Gregg, M. R. Peabody, "Effect of sample size on common item equating using the dichotomous rasch model," *Applied Measurement in Education*, vol. 33, no. 1, pp. 10–23, 2020.

[43] Q. He, C. Wheadon, "The effect of sample size on item parameter estimation for the partial credit model," *International Journal of Quantitative Research in Education*, vol. 1, no. 3, pp. 297–315, 2013.

[44] M. Antal, "On the use of elo rating for adaptive assessment," *Studia Universitatis Babes-Bolyai, Informatica*, vol. 58, no. 1, pp. 29–41, 2013.

[45] R. Pelánek, J. Rihák, J. Papoušek, "Impact of data collection on interpretation and evaluation of student models," in *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 2016, pp. 40–47.

[46] K. Taunk, S. De, S. Verma, A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 2019, pp. 1255–1260, IEEE.

[47] K. Struyven, F. Dochy, S. Janssens, "Students' perceptions about evaluation and assessment in higher education: A review," *Assessment & Evaluation in Higher Education*, vol. 30, no. 4, pp. 325–341, 2005.

[48] M. F. Rodríguez, J. Hernández Correa, M. Pérez- Sanagustín, J. A. Pertuze, C. Alario-Hoyos, "A mooc-based flipped class: Lessons learned from the orchestration perspective," in *European Conference on Massive Open Online Courses*, 2017, pp. 102–112, Springer.

[49] L. P. Prieto, Y. Dimitriadis, J. I. Asensio-Pérez, C.-K. Looi, "Orchestration in learning technology research: evaluation of a conceptual framework," *Research in Learning Technology*, vol. 23, 2015.

[50] J. Niznan, J. Papousek, R. Pelánek, "Exploring the role of small differences in predictive accuracy using simulated data," in *AIED Workshop Proceedings*, vol. 5, 2015, pp. 21–30.

Alberto Jiménez-Macías

Alberto Jiménez-Macías is a PhD student at Universidad Carlos III de Madrid. He obtained a bachelor's degree in Telematics Engineering and a master's degree in Computer Science at the Escuela Superior Politécnica del Litoral (ESPOL) (Ecuador). He carried out development and research work at the Information Technology Center (CTI-ESPOL) for 8 years. His areas of interest are Learning Analytics, Educational Data Mining and Educational Technology.

Pedro J. Muñoz-Merino

Pedro J. Muñoz-Merino is Full Professor at the Department of Telematics Engineering at Universidad Carlos III de Madrid. In 2003, he received his Telecommunication Engineering degree from the Polytechnic University of Valencia, and in 2009 his PhD in Telematics Engineering from the Universidad Carlos III de Madrid. He has been the coordinator of the LALA project, a project funded by the European Commission for the adoption of learning analytics in Latin America. He has also participated in more than 40 research projects at the international and national level, also including several contracts with companies, being the Principal Investigator in several of them related to learning analytics, educational data mining and adaptive systems. He is the co-author of more than 150 scientific publications including more than 50 in journals indexed in the JCR. In addition, he has coordinated the development and deployment of different learning analytics tools. He is also an IEEE Senior Member from 2015. His skills and experience include research and development in learning analytics, educational data mining, evaluation of learning experiences, user studies, gamification or Intelligent Tutoring System.

Carlos Delgado Kloos

Carlos Delgado Kloos received the Ph.D. degree in Computer Science from the Technische Universität München and in Telecommunications Engineering from the Universidad Politécnica de Madrid. He is Full Professor of Telematics Engineering and Rector's Delegate for Digital Microcredentials at Universidad Carlos III de Madrid, where he is also the Director of the GAST research group and Director of the UNESCO Chair on "Scalable Digital Education for All". He has carried out research stays at several universities such as Harvard, MIT, Munich, and Passau. His main research interests are in Educational Technology. He has been involved in a large number of research projects and has published around 500 articles. He has coordinated several MOOCs with over 600,000 registrations and is presently promoting the adoption of digital micro-credentials in Spain through the project CertiDigital (certidigital.es).

Reversible Image Watermarking Using Modified Quadratic Difference Expansion and Hybrid Optimization Technique

H. R. Lakshmi[1]*, Surekha Borra[2]

[1] Department of ECE, K.S. Institute of Technology, Visvesvaraya Technological University, Belagavi, Karnataka (India)
[2] Department of ECE, K. S. Institute of Technology, Bangalore, Karnataka-560109, (India)

* Corresponding author: hrl.lakshmi@gmail.com

Received 13 October 2022 | Accepted 21 January 2022 | Published 4 August 2023

Abstract

With increasing copyright violation cases, watermarking of digital images is a very popular solution for securing online media content. Since some sensitive applications require image recovery after watermark extraction, reversible watermarking is widely preferred. This article introduces a Modified Quadratic Difference Expansion (MQDE) and fractal encryption-based reversible watermarking for securing the copyrights of images. First, fractal encryption is applied to watermarks using Tromino's L-shaped theorem to improve security. In addition, Cuckoo Search-Grey Wolf Optimization (CSGWO) is enforced on the cover image to optimize block allocation for inserting an encrypted watermark such that it greatly increases its invisibility. While the developed MQDE technique helps to improve coverage and visual quality, the novel data-driven distortion control unit ensures optimal performance. The suggested approach provides the highest level of protection when retrieving the secret image and original cover image without losing the essential information, apart from improving transparency and capacity without much tradeoff. The simulation results of this approach are superior to existing methods in terms of embedding capacity. With an average PSNR of 67 dB, the method shows good imperceptibility in comparison to other schemes.

Keywords

Cuckoo Search, Difference Expansion, Fractal Encryption, Grey Wolf Optimization, Information Security, Reversible Watermarking.

DOI: 10.9781/ijimai.2023.08.002

I. Introduction

Due to the rapid usage of mobile devices and the Internet, digital images are often captured, stored, and shared on social media [1] – [3], ultimately leading to several intellectual property rights issues [4]. To address these issues, digital watermarks are frequently employed for evidence of ownership, copyright protection, and integrity verification. The key components of the digital watermarking system are embedding and extraction [5] – [7]. In the case of invisible watermarking, the watermark (owner's identification data) is invisibly embedded into all the images belonging to an owner, so that the registered watermark can be extracted from the published watermarked images solely by the owner (using his secret key and extraction algorithm) to prove the ownership in case of ownership-related disputes [8] – [9]. In invisible watermarking, it is important to embed watermarks in images without affecting the quality of the images (ex: sensitive medical images, high-quality photographs, satellite images, military maps, product designs, etc.). The distortion brought on by watermarking is often restricted to ensuring that the watermark cannot be detected, making the host (original data) and the watermarked image visually identical. Hence, the imperceptibility requirement with respect to an invisible watermark implies that the perceptual quality of the watermarked images be kept high even after watermark embedding. In watermarking, the watermark can be of different types, ranging from simple text (Owner ID, time and date stamp, company name, etc.) to binary, grayscale, or color images (depending on the quality of the company logo). Further, there may be multiple owners for the image in some applications. For example, a certain medical image can be owned by the diagnostics center, the hospital, the radiologist, and the patient. In such a scenario, the application demands a large embedding capacity. While most of the existing watermarking techniques deal with robustness in the extraction process and imperceptibility after insertion, the general requirements for digital watermarks are simplicity, high embedding capacity, and security.

Please cite this article in press as:
H. R. Lakshmi, S. Borra. Reversible Image Watermarking Using Modified Quadratic Difference Expansion and Hybrid Optimization Technique, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 141-154, 2025, http://dx.doi.org/10.9781/ijimai.2023.08.002

The transform domain approaches, which are typically complex, performed satisfactorily in terms of robustness and imperceptibility [10] – [12]. Spatial domain techniques, on the other hand, are relatively simpler, but the robustness of such schemes has fallen short. Irrespective of the embedding domain, which has been heavily used in the field of watermarking, some alterations are always applied to the cover picture when a hidden image is combined with it to generate a watermarked image. Even if they are little, these changes are undesirable in delicate applications like defense, forensics, medical imaging, etc. [13] – [14]. In reversible watermarking, the original host can be reversed from a published watermarked image after watermark extraction if required. There are many research opportunities in the field of reversible watermarks (RW), as they are suitable for applications containing sensitive images.

Least-significant-bit-based techniques are the earliest methods for reversible watermarking [15]. Contrarily, after extracting the watermark, the picture may also be reversed by modifying the coefficients of various transforms [16] – [18]. Several algorithms, including the Integer Wavelet Transform (IWT), the Riesz transform, the discrete wavelet transform (DWT), the discrete cosine transform (DCT), and the discrete Fourier transform (DFT), are used to convert the image into the frequency domain prior to embedding and extracting the watermark.

A few publications [19],[20] use singular value decomposition (SVD) to provide a strong framework against noise. The transform-domain reversible watermarking schemes discussed in the literature are proven to be more robust, but they are more complex and time-consuming [21]. While few reversible approaches require location maps [22], much research has been done to construct location maps from the host image, embed them with the watermark for later watermark extraction, and/or cover image recovery. The methodologies that use location maps are prediction error expansion (PEE) [23], histogram shifting [24], difference expansion (DE) [25],[26], integer transform [27], and hybrid techniques (combination of two or more techniques) [28],[29]. The primary goal of such strategies is to improve embedding capacity.

This article introduces a hybrid reversible image watermarking method that improves embedding capacity along with transparency and security. This study makes the following improvements to the existing reversible watermarking methods, in addition to leveraging fractal encryption to increase security:

- Hybrid soft computing using Cuckoo search and grey wolf optimization (CSGWO) is employed to determine where in the image a watermark should be embedded to ensure a higher convergence rate and visual quality.
- A novel distortion control unit is proposed to ensure imperceptibility and optimal embedding. The fitness function incorporates 3-SSIM, PSNR, cross entropy (CE) for optimal location finding in the embedding process.
- Modified quadratic difference expansion (MQDE) method to embed and extract high-capacity watermarks.

The study is organized as follows: The earlier research in the fields of difference expansion-based reversible watermarking and reversible watermarking employing optimization approaches is covered in Section II. In Section III, the preliminaries are explained. In Section IV, the suggested model is explained in detail. The experimental assessment of the suggested model is covered in Section V; Section VI provides a discussion on the performance; and Section VII provides the conclusion.

II. Literature Review

The necessity of striking a compromise between picture quality and embedding capacity in watermark methods is well demonstrated by researchers. Soft computing approaches and optimization techniques are often used to reduce the trade-off problem and to select pixel positions, blocks, or thresholds for the overall performance improvement of RW.

This study is based on difference expansion (DE)-based techniques, which Tian et al. [30] initially presented to have visual quality and satisfactory embedding capacity. A pair of pixels is chosen for one bit of secret information, based on some criteria related to the difference in their pixel values. A location map may be embedded as auxiliary data along with the secret information. Tian et al. [26] later proposed a non-compression method for masking watermark data, taking advantage of the similarity in successive pixels and increasing the value of the difference. Tzu et al. [31] proposed a lossless scheme wherein each host image pixel is split into 4-bit parts, where the nibble pairs between adjacent pixels are used for embedding the secret information. This method claims high payload capacity and full reversibility. The difference expansion method based on triplets [32] provides much more information hiding capacity, as each of the selected triplets can hide two bits of watermark. The results showed that the computational cost is comparatively less than most of the transforms, as it uses a total of only 10 additions and 6 shift operations for embedding and extraction with minimal alteration of embedding coefficients. To obtain a large payload capacity with little picture distortion, Alattar [33] explored pixel vector-based difference expansion, where a feedback system modulated the payload to be added based on the required quality. The method supports hiding data in color images and recursive embedding to increase capacity. Blocks (3×3) of the host image are classified into various categories based on their structure [34]. The variance between each block is computed, and data is embedded in each block accordingly. Secret bits and auxiliary data are embedded in separate parts of the image. Secret data is embedded by utilizing the method proposed by Alattar [33], and the auxiliary information is inserted using the substitution of LSB. Hu et al. [35] used Haar-based transforms to embed bits in both horizontal and vertical directions to increase payload capacity. A dynamic approach for pixel selection ensures balance between both embedding directions such that only the small difference values are used for embedding, thus reducing distortion.

Difference expansion (DE) schemes mostly use regions where the pixel values are similar, thus limiting the capacity. A DE-based scheme put forth by Maniriho et al. [36] has an additional mod-based function to allow for embedding pairs with both positive and negative differences. If the difference is less than 2 and greater than -2, then those pixels are chosen for embedding, thereby increasing the embedding capacity. A reduced difference expansion [37] entails selecting difference values for embedding and further reducing them before inserting payload to increase embedding capacity. Another reduced-difference expansion [38]-based scheme is put forth to increase the capacity further by embedding payload into non-changeable pixels, which in other methods were not used for embedding. Variable block size was also used in the method for capacity-based processing. Only the index of the first pixel of the variable pair is contained in the decreased size of the location map [39]. Two rounds of differential expansion are employed to increase embedding capacity.

An adaptive difference expansion (ADE) method, which typically uses a few well-known parameters for watermarking and extraction, was proposed by H.S. El-sayed et al. [40]. This technique greatly expands the embedding capacity and demonstrates the superiority of watermarked photos over many other techniques. S. Weng, et al. [41] presented an optional embedding scheme (OES) to lower the distortion based on the requirements. DE is utilized for embedding by default, and adaptive embedding fulfils the requirement of a large embedding rate with low local variance. In addition to achieving respectable performance at all embedding rates, using a local smoothness estimator and four prediction frameworks increases the number of pixels that can be embedded. Z. Zhang et al. [42] developed a quadratic difference expansion-based method for enhancing the visual quality and embedding rate. Following the first omission of the pixel points with 0 and 255 greyscale values, linear difference expansion (LDE) is used to add half of the jumbled data to the cover picture. The Quadratic Difference Expansion (QDE) is used to incorporate the remaining secret data into the previously created picture. The final watermarked picture is generated after appending the greyscale pixel values of 0 and 255. The results and simulation section showed that the technique resulted in high visual quality and embedding rates.

A computationally less expensive scheme was proposed [43] by combining downscaling and data hiding. The scheme proposed an adaptive adjustment of the capacity-quality trade-off for improved performance. It was suggested that the capacity be increased via a reversible DE technique [44]. The host image and the secret data were both encoded, and the secret was inserted in relation to a parameter representing the acceptable range of difference values. Wang [45] proposed a DE-based scheme that works bidirectionally. The pixel arrangement for data insertion is unlike other schemes, following a unique pattern. A cluster-based DE scheme [46] aimed at improving the payload capacity Upper and lower bounds are assigned, and clusters are accordingly formed. The difference expansion parameters are calculated based on the bounds and the cluster in question. Prediction error-based reversible data hiding (RDH) [47] was proposed, where data is embedded in forward and backward directions to improve capacity, and a block size 1×3 was chosen for embedding to improve the visual quality.

A few hybrid schemes used nature-inspired algorithms to optimize and improve performance parameters such as capacity and visual quality [48],[49]. The optimal brightness fitness function value was determined iteratively by the Firefly Technique (FA) [50] before secret information was embedded in the database using a DE-based algorithm. Over the earlier designs, it was observed that with this method, there was less distortion and a higher capacity. Particle Swarm Optimization (PSO) is employed for the selection of the appropriate threshold value and subsequently for the reduction of distortion in reversible watermarking based on 2D difference expansion and wavelet transform [51]. Results showed better PSNR in comparison to previous schemes. The interpolation-based expansion scheme [52] used the genetic algorithm (GA) and PSO to estimate the neighboring pixels and concluded that GA gave better results than PSO. Improvement in visual quality with respect to mean square error (MSE) and PSNR was observed [53] when embedding regions were selected using the Firefly algorithm in the DWT domain. Arnold transformation was used for watermark encryption [54] to increase the security of the reversible watermarking (RW) scheme. Both the secret image and the cover image were color images. To increase PSNR and MSE values, the strength factor is adjusted using Grey Wolf Optimization (GWO), while the secret information is incorporated using the SVD lifting procedure. The method's resilience has been successfully demonstrated for salt and pepper noise, Poisson noise, and set partitioning in hierarchical trees (SPIHT) compression.

A blind, reversible methodology that Zarrabi et al. [55] presented involves iteratively embedding the data into non-regions of interest (NROI). During embedding and extraction, ROI is identified and excluded using deep neural networks: one for segmentation and the other for classification. The resulting DCT domain scheme is lossless but not robust. A fragile reversible watermarking system based on SVD and PSO presented by Frank et al. [56] allows for dynamic capacity modification based on the desired embedding rate. Keeping the ROIs unchanged after automatic detection resulted in better quality than traditional transform domain schemes. Arsalan et al. [57] proposed using genetic programming (GP) and the integer wavelet transform (IWT) to solve the overflow and underflow problems and find the best wavelet coefficients to embed. The balance between capacity and quality is significantly reduced as companding is used for watermark insertion.

Balasamy et al. [58] developed a DWT and PSO-based method to protect medical images to find the optimal wavelet coefficients for data insertion. The approach does not produce any further data to aid in enhancing embedding capability; however, the visual quality produced is inadequate when compared to other schemes. Using Tian's DE [26], Vargas [59] created an intelligent RW scheme using a genetic algorithm (GA) to enhance the visual quality and choose the best threshold value for embedding. While simple DE is applied to 4×4, 8×8, 16×16 blocks of cover image for evaluation, the fitness function is based on MSE. Since the RW scheme is completely reversible, ROI calculations are not needed. However, embedding multiple times may lead to a smoothing effect, which is undesirable. A DWT-based RW scheme was proposed using hybrid optimization, combining two algorithms—the Tunicate Swarm Algorithm (TSA) and Simulated Annealing (SA)—for optimizing the scale factor and a deep recurrent neural network with long short-term memory (RNN-LSTM) for extraction. The authors claimed better robustness in comparison with using individual optimization schemes [60]. Ayad et al. [61] proposed a medical image watermarking approach using DWT and SVD with a text watermark encoded using QAM-16. The results showed better robustness and quality in comparison with non-hybrid schemes and were resilient against salt and pepper, and Gaussian noise. However, geometric attacks may hamper watermark detection and decoding. Kaur et al. [62] proposed a compression technique for color images using Fast Fourier Transform (FFT) compression and for optimizing 3 thresholds using the Intelligent Water Drop (IWD) algorithm: 1 for each color, by using 10 nodes for each value. The evaluation showed better SSIM values in contrast to manually chosen threshold values.

To summarize, the general measures that were taken by the researchers to improve embedding capabilities are: 1. Reduction and shrinking of the location map dimension. 2. Repeated use of DE to improve payload. 3. Make every effort to do away with the necessity for location maps. To enhance the image quality, the actions taken are: 1. Use of thresholds whose values define quality 2. Selection of the smallest (smoothest) difference-valued area of the image for embedding. However, there is an inherent trade-off between capacity and quality in the existing schemes.

This paper presents a RW technique to simultaneously meet many requirements: large embedding capacity, high security, high reversibility, high imperceptibility, and an improvement in robustness when compared to the recently published related works. The entropy value, payload, structural similarity index, and peak signal-to-noise ratio are the evaluation metrics used to track the performance of the suggested technique. While underflow and overflow are the major concerns faced by the researchers in RW schemes that degrade the system's performance, these issues are also addressed by the proposed scheme through the optimal selection of pixels using the distortion control unit.

III. Preliminaries

A. Linear Difference Expansion (LDE)

The LDE performs an integer transform on any picture pixel pair P=(s,t) to produce the difference *d* and mean *m* as stated in (1) and (2). Later, watermark bits (b) are inserted into the chosen pixel pairs of the host image.

$$d = s - t \tag{1}$$

$$m = \left[\frac{s+t}{2}\right] \tag{2}$$

The inverse transform is given in the Eqs. (5), (6).

$$s' = m + \left[\frac{d'+1}{2}\right] \tag{3}$$

$$t' = m + \left[\frac{d'}{2}\right] \tag{4}$$

The new value of difference is $d' = 2d + b$, where d is shifted to the left by one-bit b, which is the least significant bit, and is referred to as the Linear Difference Expansion (LDE). Pixel overflow results from using basic difference expansion to include hidden information or a watermark. This problem needs to be solved as the inverse transform of original pixel pairs s' and t' should fall in the range of [0, 255] for proper visibility. In addition, it is essential to limit d' as given in (5).

$$|d'| \leq (2(255 - m), 2m + 1) \tag{5}$$

B. Cuckoo Search–Grey Wolf Optimization (CSGWO)

This subsection describes how the proposed watermarking system uses the hybrid Cuckoo Search–Grey Wolf Optimization (CSGWO) algorithm to effectively select cover image pixels for watermarking while meeting the fitness function. The model combines the Grey Wolf Optimization (GWO), developed from grey wolf hunting activities [63], with a population-based algorithm called cuckoo search (CS), which uses the Levy Flight mechanism to update the new solutions (nests) in a pseudo-random manner. Thus, in the proposed method, the GWO metaheuristic incorporates CS to reinforce and increase its ability to avoid entrapment inside local optima and converge to the global minimum. The CS exploration skills are used to direct the wolves (or searching agents) to locations aided by the CS metaheuristic. In the CSGWO [64] algorithm, the GWO location update is modified to account for the CS update equation to get a faster convergence rate. The generalized equation of GWO is modified to update the position in the CSGWO algorithm, and therefore, an additional term is included in the numerator, as shown in (6) [65]:

$$\vec{G}(t+1) = \frac{\vec{G}_1+\vec{G}_2+\vec{G}_3+\vec{G}_4}{4} \tag{6}$$

Where, $\vec{G}_4$ is the position vector projected using the CS update rule and $\vec{G}_1$, $\vec{G}_2$, $\vec{G}_3$ are the hunt agents according to the best hunt agent G_α, second and third best hunt agents G_β and G_δ [66]. Cuckoo search is a metaheuristic algorithm based on the reproductive performance of cuckoos. While each egg in the nest represents a problem that is solved more effectively, this activity is utilized to update positions in the proposed CSGWO algorithm using the $\vec{G}_4$ term, defined in (7) :

$$\vec{G}_4 = \vec{G}_t + \gamma \oplus Levy\ (\lambda) \tag{7}$$

Where $\vec{G}_t$ is the agent's position in the presenter petition, γ is the step size, which attains a value from 0 to 1. $Levy\ (\lambda)$ is the Levy flight equation, which gives an arbitrary walk and is defined as $Levy \sim v = (t\text{–}\lambda)$, where λ is a constraint, whose values are in the interval [1, 3]. The addition of fourth term ($\vec{G}_4$) in proposed algorithm makes it more effective with the exploration of the search space of Levy flight.

C. Fitness Function Parameters

A gradient magnitude computation method [67] suggested that any image is composed of three regions, namely, edge, smooth, and texture, and that the segmentation is based on the threshold of the pixel gradient. Let g_{ij} represent the gradient of the original picture in (i, j) coordinates. The following guidelines form the basis for pixel categorization:

- If $g_{i,j} > TH_1$, the pixel is deliberated as edge pixel.
- If $g_{i,j} < TH_2$, pixels are processed as part of the smooth area.
- Otherwise, these pixels fall into the textured area.

Structural Similarity Index (SSIM): Since the luminous intensity of an object's surface is the result of reflection and illumination, it is preferable to eliminate the exposure effect to examine the images' structural information. The SSIM between two signals x and y is given by (8),

$$SSIM(x,y) = \frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_Y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \tag{8}$$

where μ_x and μ_y are the images x and y average intensities. σ_x and σ_y are variances, and σ_{xy} is the covariance of the two images. C_1 and C_2 are stabilizer variables that depend on the dynamic rank of the values of pixels [68].

3-SSIM: The value of 3-SSIM is calculated by comparing watermarked and original images using (9).

$$3\text{-}SSIM = A \times SSIM_{edge} + B \times SSIM_{smooth} + C \times SSIM_{texture} \tag{9}$$

where the weight factors of various regions are expressed by variables A, B and C. The notations $SSIM_{texture}$, $SSIM_{smooth}$ and $SSIM_{edge}$, indicate the SSIM values for the texture regions, smooth regions, and edge regions, respectively.

Normalized Capacity C_n: The normalized capacity of the inclusions (C_n) is calculated using (10) :

$$C_n = \frac{C_{wdc}}{C_{\frac{w}{odc}}} \tag{10}$$

where C_{wdc} and $C_{\frac{w}{odc}}$ indicate whether data can be masked with DC (distortion control) or without DC. It should be noted that the C_n falls between 0 and 1.

Cross Entropy (*CE*): Using the entropy definition, KL divergence [68], and log rules, cross entropy is defined in (11):

$$CE(p,q) = -\sum_{i=0}^{n} \quad p(x_i) log\ (q(x_i)) \tag{11}$$

where p(x) is the watermarked image and q(x) is the original image.

Peak Signal to Noise Ratio (*PSNR*) (12):

$$PSNR = 10\ log_{10}(\frac{255^2}{MSE}) \tag{12}$$

$$MSE = 1/mn \sum_{x=0}^{m-1} \quad \sum_{y=0}^{n-1} \quad [I(s,t) - d(s,t)]^2 \tag{13}$$

where, the original input image is denoted as $I(s, t)$, the recovered image is denoted as $d(s, t)$, the rows and columns of the image are denoted as m and n, and MSE denotes mean square error.

D. Fractal Algorithm

Fractal encryption is a one-to-one encryption approach that depends on modulo operations. At the first stage of the proposed watermarking approach, fractal encryption of the watermark is performed to determine the recursive contract transformation of pixels. Fractal encryption has a strong key to encrypt pictures in the context of other encryption models because of the random and disordered nature of fractals. Attributes such as zoom level, iterations, and coordinates are utilized for generating the fractal image. In this technique, two keys are generated as in (14) and (15) for encryption and decryption processes using a random number that ranges between zero and one.

$$Key\ 1 = Random \times 25 + 4 \tag{14}$$

$$Key\ 2 = Key\ 1\ \times 2 \tag{15}$$

Fig. 1 (a) and (d) show the histograms of the cover image before and after embedding the encrypted watermark image. Results indicate that the histogram of the image is consistent after watermarking. Hence, no useful statistics about the watermark can be drawn by the attacker from the published watermarked image.

Fractal encryption, when combined with the L-shaped tromino theorem, enhances the security of image transmission [69]. L shaped tromino works based on two attribute symbols, "–" or "+", and degree $\theta = 90$. The L-shaped tromino is divided into smaller trominos based on the number of iterations, determined by the size of the watermark. In Fig. 2, the first and second iterations of the L-shaped tromino are graphically depicted.

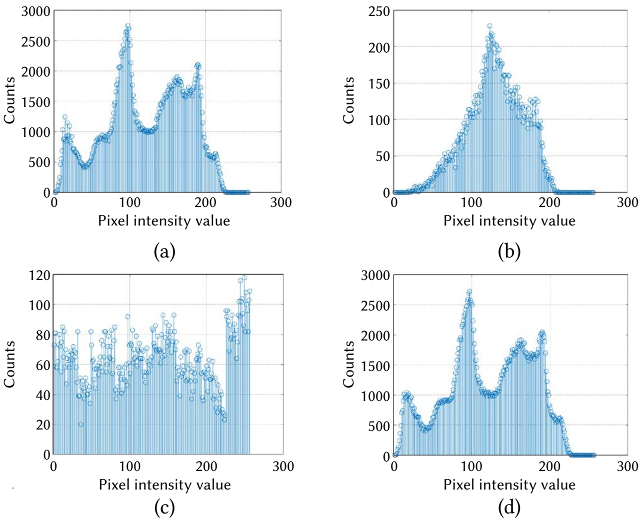


Fig. 1. Histogram graphs for (a) cover image, (b) watermark image, (c) fractal encrypted watermark and (d) watermarked cover image.

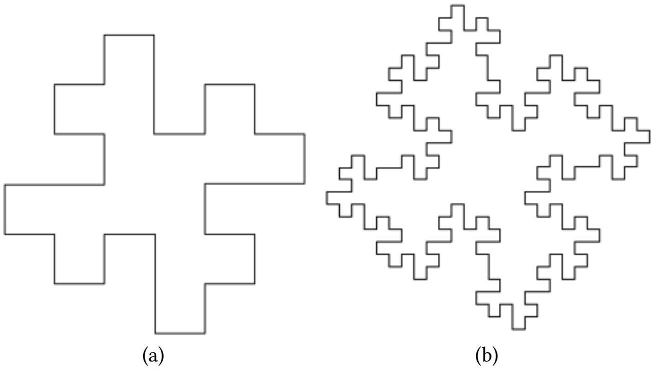


Fig. 2. (a) 1st iteration (b) 2nd iteration.

IV. Methodology

This section proposes a reversible hybrid image watermarking concept to support high-capacity (payload) secure watermarking while maintaining the visual quality of the watermarked image. This method also allows the original image to be recovered post watermark extraction. Fig. 3 shows the watermark embedding flow. The watermark image is first encrypted with the fractal encryption method to provide an extra layer of security. To safeguard data transfer with minimal system complexity, L-shaped fractal tromino-encryption is preferred. Furthermore, the best 8×8 blocks for watermark integration are selected by processing the host image using the Cuckoo search and grey wolf optimization (CSGWO) algorithm.

After exploration of the optimum location map in the host image by the CSGWO distortion control unit, modified quadratic difference expansion (MQDE) is applied for embedding the encrypted watermark to ensure better capacity and transparency. The result is a watermarked image, which can be safely published on the Internet. The embedded watermark can be extracted from the published watermarked images solely by the owner to prove ownership in cases of ownership-related disputes using his secret key and extraction algorithm.

Fig. 4 graphically shows the process of watermark extraction, where the watermarked image is initially subjected to inverse modified QDE according to the location map decided by the CSGWO distortion control unit to extract the encrypted watermark, which is later fed to a fractal decryption algorithm to finally extract the watermark and in parallel to recover the cover image from the lossy watermarked image. In the subsections that follow, each of the algorithms used in the embedding and extraction processes is discussed in detail, followed by a summary of the steps involved in applying these algorithms for the proposed reversible hybrid watermarking.

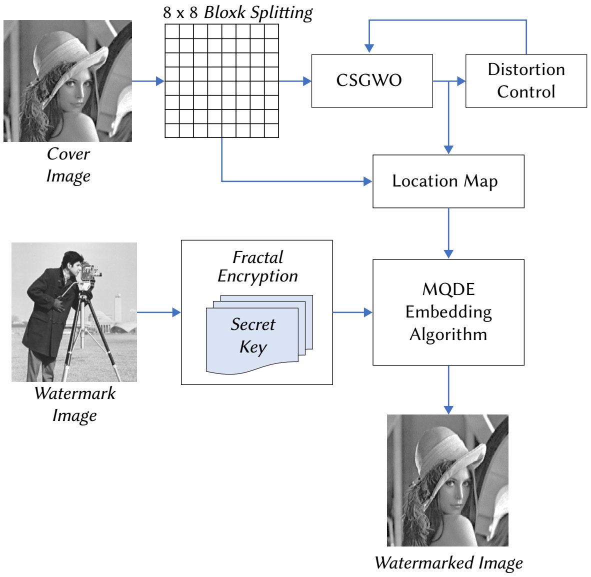


Fig.3. Watermark Embedding.

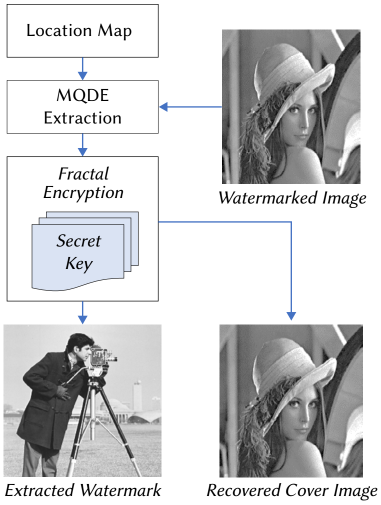


Fig.4. Watermark Extraction.

A. Modified Fitness Function

In the proposed method, the weighted sum of 3-SSIM, CE, PSNR and C_n defines the fitness function and is represented in (16)

$$F=(W_1 \times 3\text{-}SSIM)+(W_2 \times C_n)+\{W_3 \times (1-CE)\}+W_4 \times (PSNR) \quad (16)$$

The weights W_1, W_2, W_3 and W_4 must sum to 1 and are independently defined in (17) – (20)

$$W_1 = \frac{C_n \times (1-CE) \times PSNR}{3-SSIM\{C_n+(1-CE)\}+\{C_n \times (1-CE)\}+PSNR} \quad (17)$$

$$W_2 = \frac{3-SSIM \times (1-CE) \times PSNR}{3-SSIM\{C_n+(1-CE)\}+\{C_n \times (1-CE)\}+PSNR} \quad (18)$$

$$W_3 = \frac{(3-SSIM) \times C_n \times PSNR}{3-SSIM\{C_n+(1-CE)\}+\{C_n \times (1-CE)\}+PSNR} \quad (19)$$

$$W_4 = \frac{(3-SSIM) \times C_n \times (1-CE)}{3-SSIM\{C_n+(1-CE)\}+\{C_n \times (1-CE)\}+PSNR} \quad (20)$$

Replacing the valuesW_1, W_2, W_3 and W_4 in (16), the expression of the modified fitness function is represented as in (21).

$$F = \frac{4\{3-SSIM \times C_n \times (1-CE) \times PSNR\}}{3-SSIM\{C_n+(1-CE)\}+\{C_n \times (1-CE)\}+PSNR} \quad (21)$$

Starting with (14), the process continues until acceptable standard values are reached for the preferred iterations, or for 3-SSIM, PSNR, CE, and C_n.

Fig. 5 describes the steps involved in finding optimal location map using the modified fitness function obtained in the distortion control unit, to compensate for balancing hiding capacity, security, and imperceptibility in the proposed reversible watermarking.

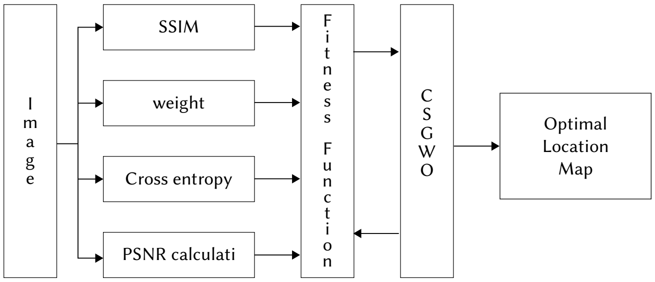


Fig.5. Distortion control unit.

B. Modified Quadratic Difference Expansion (MQDE)

The image pixel pairs s' and t' generated by difference transform in the preliminaries section using (3),(4) are utilized once again for performing quadratic watermark embedding that helps in improving the embedding capacity. The following process is mathematically stated in (1), (2), (5). d'' is the expanded difference.

$$d'' = \left[\frac{d''}{2}\right] + b \quad (22)$$

After embedding watermark using LDE, the generated watermarked image may overflow, and it returns to original image after performing modified QDE. Detailed process of modified QDE is given. Assuming the initial image pixel pair as $P = (s, t)$, the LDE process embed with the secret bit value b and the QDE process embeds the secret bit value b' using (23)-(29).

$$s' = \left[\frac{s+t}{2}\right] + \left[\ (s-t) + \frac{b+1}{2}\right] \quad (23)$$

$$t' = \left[\frac{s+t}{2}\right] - \left[(s-t) + \frac{b}{2}\right] \quad (24)$$

$$m = \left[\frac{s'+t'}{2}\right] = \left[\frac{\left[\frac{s+t}{2}\right]+\left[\frac{2(s-t)+1+b}{2}\right]+\left[\frac{s+t}{2}\right]-\left[\frac{2(s-t)+b}{2}\right]}{2}\right] \quad (25)$$

$$d = s' - t' = \left[\frac{2(s-t)+1+b}{2}\right] + \left[\frac{2(s-t)+b}{2}\right] \quad (26)$$

$$d'' = \left[\frac{\left[\frac{2(s-t)+1+b}{2}\right]+\left[\frac{2(s-t)+b}{2}\right]}{2}\right] + b' \quad (27)$$

$$s''=\left[\frac{\left[\frac{s+t}{2}\right]+\left[\frac{2(s-t)+1+b}{2}\right]+\left[\frac{s+t}{2}\right]-\left[\frac{2(s-t)+b}{2}\right]}{2}\right] + \left[\frac{\left[\left(\frac{\left[\frac{2(s-t)+1+b}{2}\right]+\left[\frac{2(s-t)+b}{2}\right]}{2}\right)\right]+b'+1}{2}\right] \quad (28)$$

$$t'' = \left[\frac{\left[\frac{s+t}{2}\right]+\left[\frac{2(s-t)+1+b}{2}\right]+\left[\frac{s+t}{2}\right]-\left[\frac{2(s-t)+b}{2}\right]}{2}\right] - \left[\frac{\left[\left(\frac{\left[\frac{2(s-t)+1+b}{2}\right]+\left[\frac{2(s-t)+b}{2}\right]}{2}\right)\right]+b'}{2}\right] \quad (29)$$

Hence, the original pixel pair (s, t) and the new pixel pair values (s'', t'') are different based on the watermark embedding value.

Consider any pixel pair (s', t') in a watermarked image (obtained through LDE) for embedding the secret information using QDE, where the to-be embedded watermark bit is represented as b'. The newly generated image pixel pairs (s'', t'') are given by (30), (31).

$$s'' = \left[\frac{s'+t'}{2}\right] + \left[\frac{\left[\frac{s'-t'}{2}\right]+b'+1}{2}\right] \quad (30)$$

$$t'' = \left[\frac{s'+t'}{2}\right] + \left[\frac{\left[\frac{s'-t'}{2}\right]+b'}{2}\right] \quad (31)$$

C. Inverse MQDE

Whenever there is a need to prove ownership, the owner first applies inverse MQDE to extract the embedded watermark from the location map [pixel pairs (s'', t'')] that is with him. Later, the original image recovery is done by applying inverse LDE.

A location map is first used to select a set of pixel pairs (s'', t'') from the watermarked image where the watermark was hidden. The average and difference values of the pixel pairs are then calculated using Equations (3), (4). Later (32) is used to normalizes the pixel values to binary 0 or 1, on applying the modulus function, which checks the pixel (s'', t'') and difference (d') values for even or odd conditions

$$A = mod(s'', 2), \;\; B = mod(t'', 2), \; C = mod(d', 2) \quad (32)$$

For instance, if the pixel value of s'' is 235, then the value is odd. The modulus operation in (32) returns A = 1. Similarly, even pixel value returns a value of 0. Based on the values of watermarked pixels and their difference, the following condition is used to extract the corresponding watermark bits.

$$Extracted\ bit = b = XNOR\ (A, B, C) \quad (33)$$

Here, XNOR performs a logical operation to check if the pixel values and difference are all even or odd. If (A, B, C) are all odd or even, the watermark bit is set to 1, otherwise it is set to 0.

The original image recovery starts with finding the LDE embedded pixel pair as given in (34) – (35)

$$s' = m' + \frac{(2(s''-t'')+1)}{2} \quad (34)$$

$$t' = m' - \frac{(2(s''-t''))}{2} \quad (35)$$

The original image pixels (s, t) are then recovered using the following equations:

$$s = \frac{6s'+2t'-2b-3}{8} \quad (36)$$

$$t = \frac{2s'+6t'+2b-1}{8} \quad (37)$$

If the secret bit that was retrieved is1, $s = s - 1$ and t is unmodified

If the secret bit that was taken is 0, s is unaltered, and t is $t - 1$.

D. Watermark Embedding Algorithm

Inputs: Secret Key, Secret Information, Cover Image

Output: Watermarked Image

1. Read the watermark and cover image.
2. Fractal encrypt the watermark using (14), (15). Convert the encrypted image into a 1-D vector.
3. Carry out CSGWO optimization on the cover picture to choose the optimal pixel pairings for embedding using the fitness function in (21) to get a higher visual quality. *I_loc* is a map that contains the locations of these pixel pairs.
4. Choose pixel pairs starting with $P = (s, t)$ based on *I_loc* (s, t).
5. Using (1,2), determine the difference and average of pixel pairs.
6. Using s (23, 24), transform the pixel pairings after LDE has expanded the difference. This results in the modified pair (s', t').
7. Find the difference and average for the pair (s', t') using (25, 26).
8. Expand the difference using (27) and use (30,31), to determine the final MQDE converted pair (s'', t'').
9. For each pair of pixels from *I_loc*, repeat steps 4 through 8 to create a watermarked picture.

E. Watermark and Cover Image Recovery Algorithm

Inputs: Location Map and Watermarked Image

Outputs: Expected Watermark, Recovered Cover Image

1. The selection of pixel pairs in the watermarked image P= (s'', t'') with the help of *I_loc*,
2. Utilizing (1,2), discover the average value and difference of pixels.
3. Utilizing (32), check for odd and even circumstances of difference and pixel pair.
4. Utilizing (33), discover the extracted watermark bit.
5. In two steps, recover the cover picture. Using (34), first separate the LDE changed pixels (s', t') from (s', t'') (35). Then, using (36), reconstruct the cover image pixel pair (s, t) (37).
6. Modify the anticipated pixel values for the cover image according to the recovered watermark bit in step 4.
7. For each pair of *I_loc* pixel pairs, repeat steps 1-6 to recover the cover image.

V. Experimental Results

A. Performance of Watermarking in the Absence of Attacks

This section presents the quantitative and qualitative findings of the suggested model in the absence of attacks on published watermarked images. In Fig. 6, the subjective results of the embedding and extraction processes are shown, considering the greyscale Baboon image (512 x 512) as the original image (to be watermarked) and the cameraman images of two sizes (128 x 128 and 32 x 32) as the watermarks. These results after embedding the individual watermarks exhibits high imperceptibility of the watermark in watermarked images along with high-quality reversed images after watermark extraction, regardless of the size of the watermark (either 128 × 128 or 32 × 32). In the absence of an attack on the watermarked picture, Table I tabulates the objective findings after comparing the similarity between the original image and recovered original image, original image, and watermarked image, and original and extracted watermarks.

When a 128× 128 greyscale pixels (131072 bits) watermark is embedded in a 512×512 greyscale cover image, the suggested technique derived an average of 46 dB PSNR when evaluated on original and watermarked images, indicating high imperceptibility. An average of 67 dB PSNR was noticed, when calculated between the original and recovered cover images, indicating high reversibility. When the watermark size is reduced to 32×32 pixels, the imperceptibility and reversibility are even better. It is to be noted that the watermark is extracted without any loss in the absence of attacks, as PSNR (OW, EW) is infinity, where EW is the extracted watermark and OW is the original watermark.

B. Time Complexity Analysis

Fig. 7 depicts a comparison of seven optimization strategies in terms of the calculation time necessary to get the ideal value using different reference functions with varying numbers of design variables. The calculation time of seven optimization approaches is accounted for by the NFE (number of function evaluations) in the collection of reference functions [70]. It was observed that the GWO and CSA take the least amount of computing time. Thus, the hybrid CSGWO approach offers the least computation time among other hybrid methods. The authors of [71] also concluded that GWO performs better than other metaheuristic algorithms in terms of complexity.

C. Embedding Capacity Analysis

Fig. 8–12 illustrates the impact of changes in embedding capacity on various watermarking evaluation parameters. Embedding

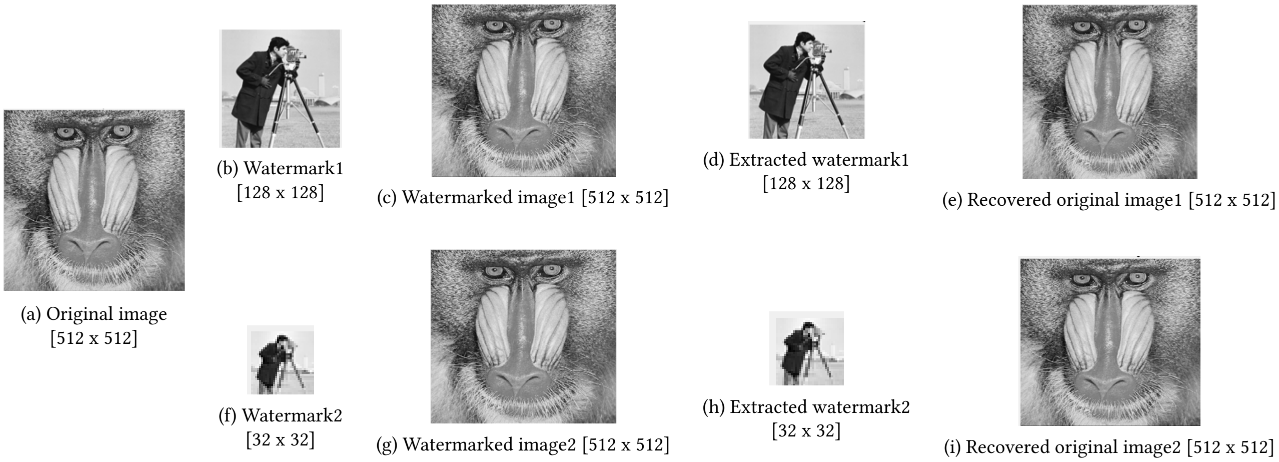

(a) Original image [512 x 512]
(b) Watermark1 [128 x 128]
(c) Watermarked image1 [512 x 512]
(d) Extracted watermark1 [128 x 128]
(e) Recovered original image1 [512 x 512]
(f) Watermark2 [32 x 32]
(g) Watermarked image2 [512 x 512]
(h) Extracted watermark2 [32 x 32]
(i) Recovered original image2 [512 x 512]

Fig. 6. Subjective results of proposed reversible watermarking.

TABLE I. Objective Results in the Absence of Attacks

Cover Image	PSNR (OI, WI)	SSIM (OI, WI)	PSNR (OI, ROI)	SSIM (OI, ROI)	PSNR (OW, EW)	SSIM (OW, EW)
	For watermark of size 128×128					
Baboon	46.384	0.99442	65.56	0.99987	Inf	1
Lena	46.357	0.97918	68.10	0.99988	Inf	1
Peppers	46.359	0.98308	67.94	0.99988	Inf	1
Barbara	46.359	0.98567	67.93	0.99992	Inf	1
	For watermark of size 32×32					
Baboon	51.601	0.99894	71.95	1	Inf	1
Lena	51.597	0.99494	72.37	0.99988	Inf	1
Peppers	51.589	0.99509	73.36	0.99997	Inf	1
Barbara	51.597	0.99498	72.38	0.99998	Inf	1
OI – Original Image, WI – Watermarked Image, ROI – Recovered Original Image, OW- Original Watermark, EW- Extracted Watermark						

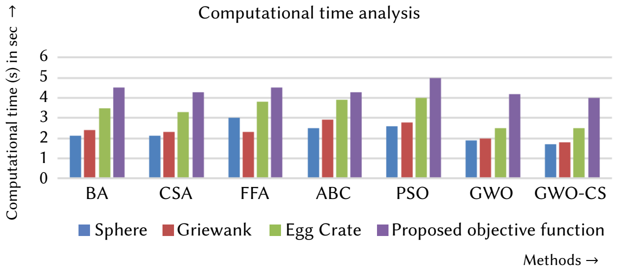


Fig.7. Computational time analysis for various optimization methods.

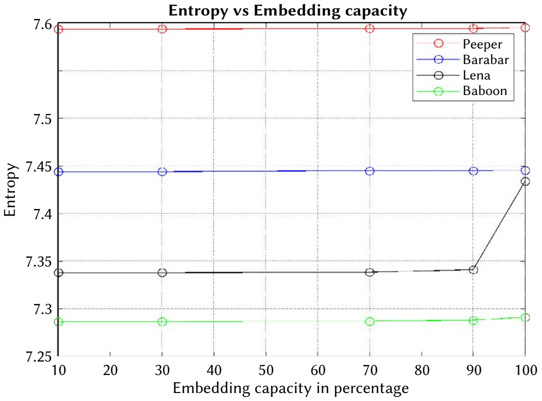


Fig. 8. Embedding capacity vs. Entropy.

capacity has been changed with respect to 10, 30, 70, 90, and 100 percentages. Fig. 8 represents the outcome between entropy and embedding capacity, where entropy is calculated among different images. It is to be noted that there is not much deviation in entropy as embedding capacity changes. Fig. 9 shows the SSIM vs. embedding capacity, and it can be clearly observed that lower embedding rates lead to less distortion in the watermarked image. Similarly, Fig. 10 represents PSNR vs. embedding capacity, where PSNR values go down as embedding capacity increases. It is interesting to note that even with 100% embedding capacity, the PSNR values remain above 50 dB. Furthermore, Fig. 11 shows that NCC values decrease as embedding capacity increases but remain significantly good even at 100% capacity. Fig. 12 indicates that MSE values go up with capacity.

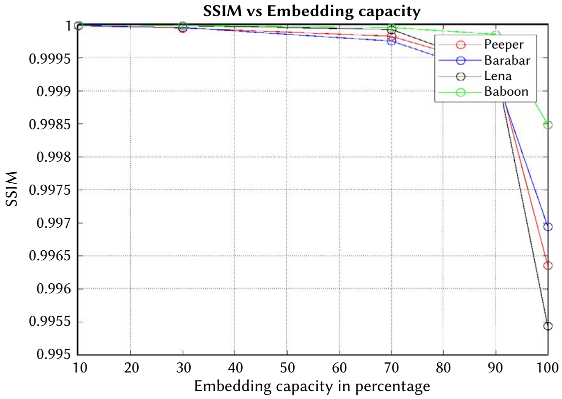


Fig. 9. Embedding capacity vs. SSIM.

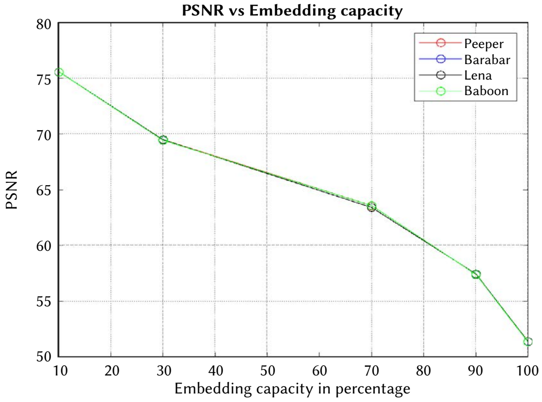


Fig. 10. Embedding capacity vs. PSNR.

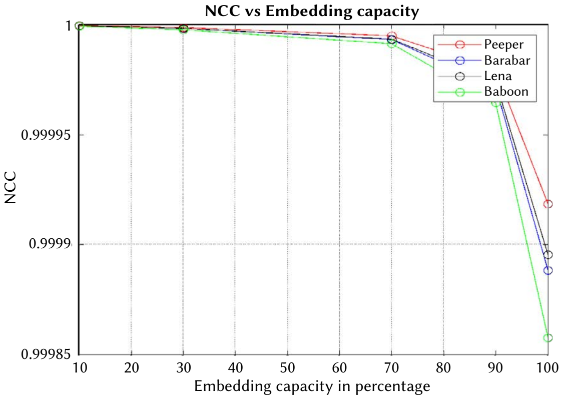


Fig. 11. Embedding capacity vs. NCC.

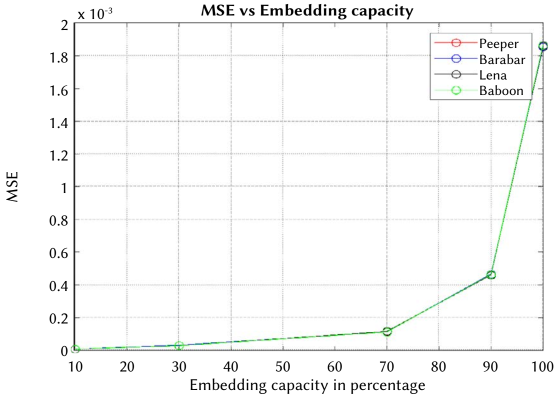


Fig. 12. Embedding capacity vs. MSE.

D. Comparison of Hybrid CSGWO With GWO

With Table II listing the simulation parameters for the proposed CSGWO, it can be concluded from Fig.13, that the convergence rate is high in the proposed CSGWO as compared to simple GWO. At the time of initial iteration, best-score achieved by hybrid approach is better than GWO. As iteration increases the obtained best score in GWO and hybrid CSGWO become similar.

TABLE II. Simulation Parameters for the Proposed CSGWO

Number of search agents	40
Maximum Iteration	100
Search Dimension	2
Domain of Search for alpha	[-1000 1000]
Domain of search for beta	[-1 1]
Domain of search for delta	[0 1000]

Table III depicts the comparative performance analysis of Grey Wolf (GWO), Cuckoo Search (CS) algorithm and Hybrid (CSGWO) approaches. The PSNR and SSIM when calculated for various benchmark images after embedding a payload of 16384 bytes (128×128) indicates that the CSGWO gives better results. Approximately 2 dB improvement is achieved with the proposed hybrid CSGWO as compared to traditional GWO and CS respectively.

E. Attack Resilience of Watermarking

Attacks [72], [73] that are both purposeful and unintended are considered while evaluating the suggested watermarking strategy. Considering Cameraman (128×128) as a watermark, Baboon (256×256) as an original image, the PSNR calculated between the original and extracted watermarks in the presence of a few attacks is tabulated in Table 4. Results show that the method is resistant to histogram equalization, cropping, and salt and pepper noise.

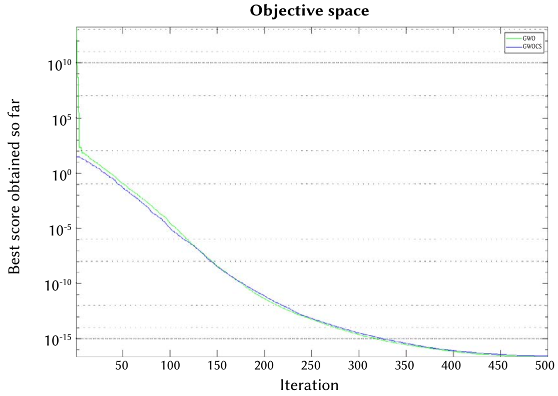


Fig. 13. Convergence graph of GWO and hybrid CSGWO.

TABLE IV. Performance in the Presence of Attacks

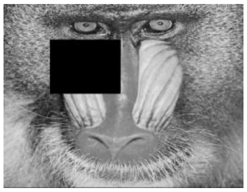
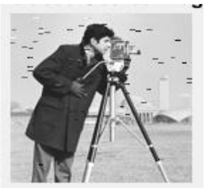
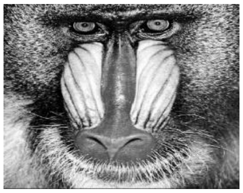
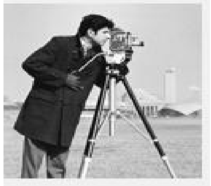
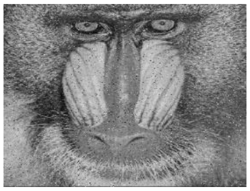
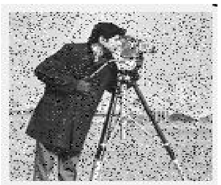
Attacks	Watermarked image	Extracted watermark
10% Crop		
PSNR (OI, ROI)	26.971	
PSNR (OW, EW)	26.786	
Histogram equalization		
PSNR (OI, RI)	17.619	
PSNR (OW, EW)	Inf	
Salt and Pepper noise (0.05 density)		
PSNR (OI, RI)	18.58	
PSNR (OW, EW)	16.61	

TABLE III. Performance Evaluation of Reversible Watermarking With Respect to Various Optimization Techniques

Cover Image	PSNR (CSGWO)	SSIM (CSGWO)	PSNR (GWO)	SSIM (GWO)	PSNR (CS)	SSIM (CS)
	For watermark of size 128×128					
Baboon	46.384	0.99442	44.845	0.9895	44.163	0.9820
Lena	46.357	0.97918	44.894	0.9613	44.920	0.9728
Peppers	46.359	0.98308	43.269	0.9598	44.249	0.9609
Barbara	46.359	0.98567	44.738	0.9730	44.209	0.9789

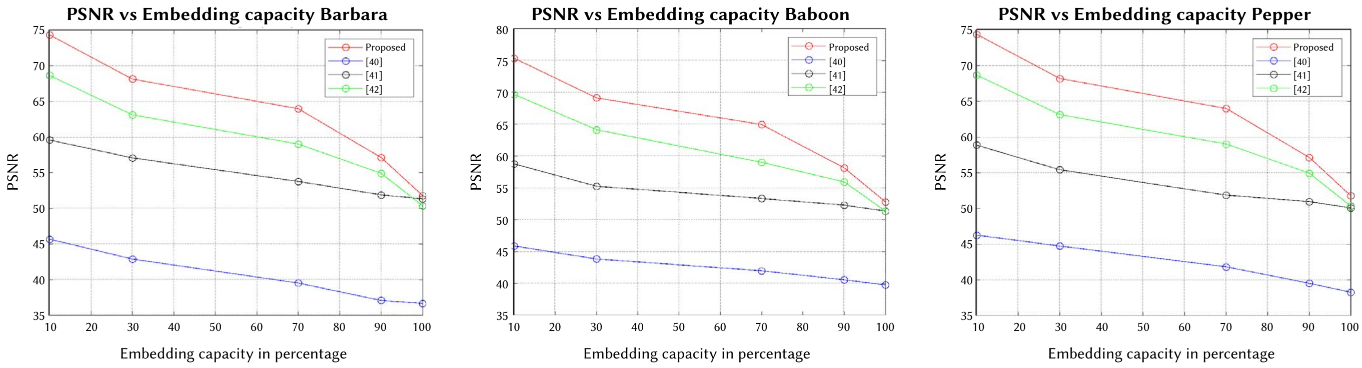


Fig. 14. Proposed method vs research works [40], [41] and [42].

TABLE V. Comparative Analysis With State-of-the Art Techniques

	[43]		[44]		[45]		[46]		[47]		Proposed	
	Capacity	PSNR	Capacity	PSNR	Capacity	PSNR	Capacity	PSNR	Capacity	PSNR	Capacity	PSNR
Airplane	228863	25.97	164962	32.98	-	-	71,680	53.6	51376	52.60	259422	55.2
Baboon	459737	20.53	128256	30.25	26000	32	71,680	53.5	16011	51.47	258676	55.38
Barbara	290234	23.43	-	-	31000	32	-	-	30559	51.80	261226	56.36
Boat	301995	25.26	142506	30.89	-	-	71,680	53.87	29686	51.73	257662	55.61
Couple	299203	24.02	-	-	-	-	-	-	-	-	258894	55.42
Lena	224528	28.35	172627	32.28	182000	32	-	-	36607	52.02	261298	55.35
Peppers	223295	26.79	181258	33.3	104000	32	-	-	34797	51.88	258458	55.36

VI. Comparative Investigation and Discussion

The embedding capacity (watermark size with respect to the original image) is varied in terms of percentages (10%, 30%, 70%, 90%, and 100%) and the corresponding PSNR calculated between the original and watermarked images is compared in Fig. 14 for three different reversible watermarking schemes proposed in the literature. The results indicate that the proposed method outperforms [40]–[42] and exhibits high imperceptibility (high PSNR) under different embedding capacities for four benchmark images. Table 5 extends the comparative analysis of other state-of-the art techniques. The proposed method supports large-sized watermarks while still maintaining high invisibility. Hence, it is concluded that the proposed hybrid algorithm mainly reduces the capacity-invisibility tradeoff compared to the related invisible reversible watermarking schemes.

An analytical comparison of the suggested method with the current hybrid and non-hybrid models is shown in Table 6. Adaptive Difference Expansion (ADE)-based techniques claim to improve embedding capacity at the expense of imperceptibility. Further, the method is complex as it requires many parameters, such as the stego-image, average, difference, maximum, and minimum pixel values of surrounding pixels, apart from the watermarked image and location maps for watermark extraction and the watermark recovery process. On the same Baboon grayscale image of size 512×512, the optional embedding scheme (OES) claims high imperceptibility (53 dB PSNR) but can only embed a 15000-bit watermark. The hybrid models, which combined LDE and QDE, claimed high imperceptibility after reporting a PSNR of 76.87 dB. An additional layer of protection, encryption, is applied to the watermark and has the benefit of not requiring a map of its position. However, it has an additional process of first removing 0 and 255 valued pixels and then again attaching them for both embedding and extraction. The performance of these three methods was not evaluated and reported for complexity, robustness, and reversibility. Furthermore, there is no optimization of pixel selection to improve visual quality. The hybrid models that combined difference expansion with a genetic algorithm for embedding watermark bits supported a maximum capacity of 0.7 bpp for Lena and Boats images at 32.43 dB and 31.3 dB PSNR, respectively. The imperceptibility of this hybrid model is low, and the smoothing effect [26] remains. It has been shown that when capacity rises, visual quality falls, creating a trade-off. To control quality deterioration, a high-capacity RW system must be designed. A favorable capacity-quality ratio is largely maintained by our proposed approach, which could embed 128×128 greyscale image (131072 bits) and could achieve a high PSNR and high visual quality based on the results presented in Table VI. Hence, compared to relevant algorithms, the approach proposed in this paper exhibits high embedding capacity, supports grayscale images as watermarks, and retains the quality of both watermarked and reversed images (an average PSNR of 67 dB). The suggested technique is resistant to many common attacks.

The technique is further strengthened by using fractal encryption of the watermark before its embedding. The secret key used in the fractal encryption method is neither being transferred over the network nor embedded into the image. It is only held by the owner and is used by him for encryption (before watermark embedding) and decryption (after watermark extraction). For the sake of secret key recovery, an attacker could attempt to distinguish any notable information between the normal image and its encryption version. An image with considerable visual information is distinguished by strong correlation and redundancy between surrounding pixels, whether in vertical, diagonal, or horizontal orientations. A well-designed encryption algorithm [74] – [76] should be capable of concealing such links between neighboring pixels while demonstrating zero correlation. The number of pixel change rate (NPCR) parameter is calculated and compared with various encryption algorithms employed in earlier state-of-the-art reversible watermarking techniques to estimate the correlation performance of fractal encryption for its applicability in the proposed watermarking. Because the achieved NPCR (99.65) was close to the theoretical value of 100, fractal encryption was chosen.

TABLE VI. Existing Models Comparison

	Non-Hybrid Techniques		Hybrid Techniques		
	[40]	[41]	[42]	[59]	**Proposed**
Methodology	Adaptive Difference Expansion (ADE)	Optional Embedding Scheme (OES)	Quadratic Difference Expansion (QDE)	Difference Expansion with Genetic Algorithm	Hybrid Modified Difference with Fractal encryption and GWO
Watermark Type	Bit stream	Bit stream	Binary image	Bit stream	Greyscale image
Secret keys	No	No	Required	No	Required
Optimization for selection of pixels	No	No	No	Genetic Algorithm	Grey Wolf Optimization
Location map	No	Required	No	Required	Required
Inputs of Watermark embedding phase	Cover Image, Embedding parameters ep1, ep2, ep3, Watermark bit stream; Average, Maximum and minimum of pixel values of surrounding pixels	Cover Image, Location map, Overhead information (EC), Watermark bitstream	Cover Image with 0 and 255 pixels removed, Scrambled Watermark	Cover Image, Location map, Watermark bitstream	Cover image, Encrypted watermark, Location map
Watermark Size	88448 bits	15000 bits	1024 bits	183500 bits	131072 bits
PSNR (OI, WI)	39.76 dB	53 dB	76.87 dB	32.43 dB	66.38 dB

The proposed hybrid model combined meta heuristics to choose optimal locations for embedding watermark bits, thereby improving performance by reducing tradeoffs between embedding capacity and imperceptibility. The disadvantage of a particular optimization algorithm can be overcome by utilizing a complementary feature of another algorithm, thus gaining from both algorithms, and resulting in better performance. GWO was selected in the proposed method over other meta-heuristics because of its relatively simple structure and lower storage requirements. Only two parameters needed to be tuned, and the decision variables were also limited. GWO aims to find the individual with the best fitness value, thus limiting the global search and increasing the chance of encountering a local optimum. However, the update mechanism has two major drawbacks in optimizing real-world functions: first, due to the use of the best global solutions found so far, the algorithm converges very quickly to a local optimal solution and loses its optimization power significantly; second, it causes the loss of a variety of new populations in each iteration of the algorithm. To fix these two shortcomings and strengthen the GWO algorithm, CS is incorporated into GWO for better performance regarding good exploration. It is much easier to jump from the current region to another, as CS updates the nest's positions with a certain probability independent of the search path, and with random directions. In CSGWO, the GWO location update is modified to account for the CS update equation to get a faster convergence rate in comparison to GWO.

The hybrid approach of combining meta-heuristic algorithms may also lead to an additional computational cost. To make a data hiding algorithm reversible or lossless, it is desirable to keep the complexity low, which directly impacts the robustness. Since the proposed scheme uses a spatial domain technique, its robustness is limited to a few attacks. It is still difficult to create a spatial domain RW scheme that is resistant to all types of attacks. Hiding multiple secret images or watermarks using the same scheme to hide more information without a tradeoff in imperceptibility is another challenge to work toward. Further, the performance of a hybrid meta-heuristic approach depends on the defined fitness function. Therefore, one cannot generalize that a hybrid approach always outperforms individual meta-heuristic algorithms. Thus, choosing a hybrid approach suitable for the given fitness function remains another challenge, which may be taken up as future work.

VII. Conclusion

This study introduced a reversible watermarking technique based on MQDE with hybrid optimization and fractal encryption, to meet the three important properties: imperceptibility, embedding capacity, and security. Correlated to former embedding techniques, CSGWO with the MQDE method considerably expands the optimal visual quality and embedding capacity. A novel data-driven distortion control unit is used for defining the optimization parameters with each iteration. The proposed model achieved satisfactory imperceptibility and was observed to be superior to the existing models (ADE, OES, and QDE) considering robustness against salt and pepper noise, cropping attacks, and histogram equalization. An average PSNR of 67 dB was achieved. The L-shaped tromino method combined with fractal encryption produces a protected image watermark. The suggested algorithms exceed the current methodologies with all these benefits, especially in terms of embedding capacity and reversibility, without sacrificing invisibility, and with the ability to withstand minimal attacks. Future research focuses on improving the reversibility performance of the proposed system in the presence of geometric attacks. While the method can easily be extended to be compatible with color images, the development of methods for the elimination of location maps can be explored. More meta-heuristics and combinations of them can be explored to determine the best approach for the given application.

Conflict of Interest

Author 1 (Lakshmi H R) declares that she has no conflict of interest.

Author 2 (Surekha Borra) declares that she has no conflict of interest.

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Funding

Nil.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] H. R. Lakshmi and S. Borra, "Difference expansion based reversible watermarking algorithms for copyright protection of images: state-of-the-art and challenges," *International Journal of Speech Technology*, vol. 24, no. 24, pp. 823-852, 2021, doi: https://doi.org/10.1007/s10772-021-09818-y.

[2] K. Curran and R. Lautman, "The Problems of Jurisdiction on the Internet," *International Journal of Ambient Computing and Intelligence (IJACI)*, vol. 3, no. 3, pp. 36-42, 2011, doi: https://doi.org/10.4018/jaci.2011070105.

[3] D. Quinn, L. Chen, and M. Mulvenna, "Social network analysis: A survey," *International Journal of Ambient Computing and Intelligence (IJACI)*, vol. 4, no. 3, pp. 46-58, 2012, doi: https://doi.org/10.4018/jaci.2012070104.

[4] S. Borra and H. R. Lakshmi, "Visual Cryptography Based Lossless Watermarking for Sensitive Images," in *International Conference on Swarm, Evolutionary, and Memetic Computing*, vol. 9873, B. Panigrahi, P. Suganthan, S. Das and S. Satapathy, Eds. Cham: Springer International Publishing, 2015, pp. 29-39.

[5] D. Ariatmanto and F. Ernawan, "Adaptive scaling factors based on the impact of selected DCT coefficients for image watermarking," *Journal of King Saud University-Computer and Information Sciences,* vol. 34, no. 3, pp. 605-614, 2020, doi: https://doi.org/10.1016/j.jksuci.2020.02.005.

[6] S. Borra Surekha, H. R. Lakshmi, N. Dey, A. S. Ashour, and F. Shi, "Digital Image Watermarking Tools: State-of-the-Art," in *2nd International Conference on Information Technology and Intelligent Transportation Systems,* vol. 296, V. E. Balas, C. J. Lakhmi, X. Zhao, F. Shi, Frontiers in Artificial Intelligence and Applications: IOS Press, 2017, pp. 450-459.

[7] S. Borra, R. Thanki, and N. Dey, *Digital image watermarking: theoretical and computational advances*, New York, USA: CRC Press, 2018.

[8] Y. Zhang and Y. Sun, "An image watermarking method based on visual saliency and contourlet transform," *Optik,* vol. 186, pp. 379-389, 2019, doi: https://doi.org/10.1016/j.ijleo.2019.04.091

[9] H. R. Lakshmi, B. Surekha. and S. Viswanadha Raju, "Real-time Implementation of Reversible Watermarking," *In: Intelligent Techniques in Signal Processing for Multimedia Security*, vol. 660, N. Dey, V. Santhi, Eds. Cham: Springer International Publishing, 2017, pp. 113-132.

[10] H.J. Ko, C.T. Huang, G. Horng and W.A.N.G. Shiuh-Jeng, "Robust and blind image watermarking in DCT domain using inter-block coefficient correlation," *Information Sciences,* vol. 517, pp. 128-147, 2020, doi: https://doi.org/10.1016/j.ins.2019.11.005 517.

[11] B. Surekha, G. Swamy and K. S. Rao, "A multiple watermarking technique for images based on visual cryptography," *International Journal of Computer Applications,* vol. 1, no. 11, pp. 77-81, 2010, doi: 10.5120/236-390

[12] A. K. Pal, P. Das and N. Dey, "Odd-even embedding scheme based modified reversible watermarking technique using Blueprint", arXiv preprint, arXiv:1303.5972, 2013. Accessed: Oct. 12, 2022. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1303/1303.5972.pdf

[13] B. Surekha and G. Swamy, "A semi-blind image watermarking based on Discrete Wavelet Transform and Secret Sharing," in *2012 IEEE International Conference on Communication, Information & Computing Technology (ICCICT),* Mumbai, India, 2012, pp. 1-5.

[14] M. Cinque, A. Coronato and A. Testa, "Dependable Services for Mobile Health Monitoring Systems," *International Journal of Ambient Computing and Intelligence*, vol. 4, no.1, pp. 1-15, 2012, doi: https://doi.org/10.4018/jaci.2012010101

[15] S. L. Li, K. C. Leung, L. M. Cheng and C. K. Chan, "Data Hiding in Images by Adaptive LSB Substitution Based on the Pixel-Value Differencing," in *First IEEE International Conference on Innovative Computing, Information and Control - Volume I (ICICIC'06)*, Beijing, China, 2006, pp. 58-61.

[16] S. H. Wang and Y. P. Lin, "Wavelet tree quantization for copyright protection watermarking," in *IEEE Transactions on Image Processing*, vol. 13, no. 2, pp. 154-165, 2004, doi: 10.1109/TIP.2004.823822.

[17] W. H. Lin, Y. R. Wang, S. J. Horng, T. W. Kao and Y. Pan, "A blind watermarking method using maximum wavelet coefficient quantization," *Expert Systems with Applications*, vol. 36, no. 9, pp. 11509-11516, 2009, doi: https://doi.org/10.1016/j.eswa.2009.03.060

[18] A. A. Reddy and B. N. Chatterji, "A new wavelet-based logo-watermarking scheme," *Pattern Recognition Letters,* vol. 26, no. 7, pp. 1019-1027, 2005, https://doi.org/10.1016/j.patrec.2004.09.047

[19] M. Yamni, A. Daoui, H. Karmouni, M. Sayyouri, H. Qjidaa and J. Flusser, "Fractional Charlier moments for image reconstruction and image watermarking," *Signal Processing*, vol. 171, pp. 107509, 2020, doi: https://doi.org/10.1016/j.sigpro.2020.107509

[20] S. Chakraborty, S. Chatterjee, N. Dey, A.S. Ashour and A.E. Hassanien, "Comparative Approach Between Singular Value Decomposition and Randomized Singular Value Decomposition-based Watermarking," *In Intelligent Techniques in Signal Processing for Multimedia Security,* vol. 660, N. Dey, V. Santhi, Eds. Chams: Springer International Publishing, 2017, pp. 133-149. vol 660.

[21] H. R. Lakshmi and B. Surekha, "Asynchronous Implementation of Reversible Image Watermarking Using Mousetrap Pipelining," in *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, Bhimavaram, India, 2016, pp. 529-533.

[22] S. Gujjunoori and B. B. Amberker, "DCT based reversible data embedding for MPEG-4 video using HVS characteristics," *Journal of information security and applications*, vol. 18, no. 4, pp. 157-166, 2013, https://doi.org/10.1016/j.istr.2013.01.002.

[23] M. Liu, H. S. Seah, C. Zhu, W. Lin and F. Tian, "Reducing location map in prediction-based difference expansion for reversible image data embedding," *Signal Processing*, vol. 92, no. 3, pp. 819-828, 2012, doi: https://doi.org/10.1016/j.sigpro.2011.09.028.

[24] Z. Ni, Y. Q. Shi, N. Ansari and W. Su, "Reversible data hiding," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 3, pp. 354-362, 2006, doi: 10.1109/TCSVT.2006.869964.

[25] C. C. Chang and T. C. Lu, "A difference expansion-oriented data hiding scheme for restoring the original host images," *Journal of Systems and Software*, vol. 79, no. 12, pp. 1754-1766, 2006, https://doi.org/10.1016/j.jss.2006.03.035.

[26] J. Tian, "Reversible data embedding using a difference expansion," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 8, pp. 890-896, 2003, doi: 10.1109/TCSVT.2003.815962.

[27] X. Wang, X. Li, B. Yang and Z. Guo, "Efficient Generalized Integer Transform for Reversible Watermarking," *in IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 567-570, 2010, doi: 10.1109/LSP.2010.2046930.

[28] H. C. Huang, F. C. Chang and W. C. Fang, "Reversible data hiding with histogram-based difference expansion for QR code applications," in *IEEE Transactions on Consumer Electronics*, vol. 57, no. 2, pp. 779-787, 2011, doi: 10.1109/TCE.2011.5955222.

[29] W. L. Tai, C. M. Yeh and C. C. Chang, "Reversible Data Hiding Based on Histogram Modification of Pixel Differences," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 6, pp. pp. 906-910, 2009, doi: 10.1109/TCSVT.2009.2017409.

[30] J. Tian. "Reversible watermarking by difference expansion," in *Proceedings of workshop on multimedia and security*, vol. 19, J. Dittmann, J. Fridrich, P. Wohlmacher, Eds. ACM, 2002, Juan-les-Pins: ACM, 2002. Multimedia and Security Workshop at ACM Multimedia '02, December 6, 2002, pp. 19–22.

[31] T. C. Lu, and C. C. Chang, "Lossless nibbled data embedding scheme based on difference expansion," *Image and Vision Computing*, vol. 26, no. 5, pp. 632-638, 2008, doi: https://doi.org/10.1016/j.imavis.2007.07.011.

[32] E. Chrysochos, V. Fotopoulos and A. N. Skodras, "A new difference expansion transform in triplets for reversible data hiding," *International Journal of Computer Mathematics*, vol. 88, no. 10, pp. 2016-2025, 2011, doi: https://doi.org/10.1080/00207160.2010.539210

[33] A. M. Alattar, "Reversible watermark using the difference expansion of a generalized integer transform," in *IEEE Transactions on Image Processing*, vol. 13, no. 8, pp. 1147-1156, 2004, doi: 10.1109/TIP.2004.828418.

[34] J. Y. Hsiao, K. F. Chan and J. M. Chang, "Block-based reversible data embedding," *Signal Processing*, vol. 89, no. 4, pp. 556-569, 2009, doi: https://doi.org/10.1016/j.sigpro.2008.10.018

[35] Y. Hu, H. K. Lee, K. Chen and J. Li, "Difference Expansion Based Reversible Data Hiding Using Two Embedding Directions," in *IEEE Transactions on Multimedia*, vol. 10, no. 8, pp. 1500-1512, 2008, doi:10.1109/TMM.2008.2007341.

[36] P. Maniriho and T. Ahmad, "Information hiding scheme for digital images using difference expansion and modulus function," *Journal of king saud university-computer and information sciences,* vol. 31, no. 3, pp. 335-347, 2019, doi: https://doi.org/10.1016/j.jksuci.2018.01.011

[37] P. Maniriho and T. Ahmad, "Enhancing the Capability of Data Hiding Method Based on Reduced Difference Expansion," *Engineering Letters*, vol. 26, no. 1, pp. 45-55, 2018.

[38] M. H. A. Al-Hooti, T. Ahmad and S. Djanali, "Improving the Capability of Reduced Difference Expansion based Digital Image Data Hiding," *IAENG International Journal of Computer Science*, vol. 46, no. 4, 2019.

[39] S. Gujjunoori and M. Oruganti, "Difference expansion based reversible data embedding and edge detection," *Multimedia Tools and Applications*, vol. 78, pp. 25889–25917, 2019, doi: https://doi.org/10.1007/s11042-019-07767-y.

[40] H. S. El-sayed, S. F. El-Zoghdy and O. S. Faragallah, "Adaptive difference expansion-based reversible data hiding scheme for digital images," *Arabian Journal for Science and Engineering*, vol. 41, no. 3, pp. 1091-1107, 2016, doi: https://doi.org/10.1007/s13369-015-1956-7

[41] S. Weng, J. S. Pan and L. Zhou, "Reversible data hiding based on the local smoothness estimator and optional embedding strategy in four prediction modes," *Multimedia Tools and Applications*, vol. 76, no. 11, pp. 13173-13195, 2017, doi: https://doi.org/10.1007/s11042-016-3693-7.

[42] Z. Zhang, M. Zhang, L. Wang, "Reversible Image Watermarking Algorithm Based on Quadratic Difference Expansion," *Mathematical Problems in Engineering*, vol. 2020, 2020. https://doi.org/10.1155/2020/1806024.

[43] A. A. Mohammad, A. Al-Haj and M. Farfoura, "An improved capacity data hiding technique based on image interpolation," *Multimedia Tools and Applications*, vol. 78, no. 6, pp. 7181-7205, 2019, doi: https://doi.org/10.1007/s11042-018-6465-8

[44] R. Anushiadevi, P. Praveenkumar, J. B. B. Rayappan and R. Amirtharajan, "Reversible data hiding method based on pixel expansion and homomorphic encryption," *Journal of Intelligent & Fuzzy Systems*, vol. 39, no. 3, pp. 2977-2990, 2020, doi: 10.3233/JIFS-191478.

[45] W. Wang, "A reversible data hiding algorithm based on bidirectional difference expansion," *Multimedia Tools and Applications*, vol. 79, no. 9, pp. 5965-5988, 2020, doi: https://doi.org/10.1007/s11042-019-08255-z

[46] A. J. Ilham, and T. Ahmad, "Reversible Data Hiding Scheme based on General Difference Expansion Cluster," *International. Journal of Advance Soft Computing and Applications*, vol. 12, no. 3, pp. 11-24, 2020.

[47] M. Abdul Wahed and H. Nyeem, "Reversible data hiding with dual pixel-value-ordering and minimum prediction error expansion," *Plos one*, vol. 17, no. 8, 2022, doi: https://doi.org/10.1371/journal.pone.0271507

[48] N. Dey, J. Chaki, L. Moraru, S. Fong and X. S. Yang, "Firefly Algorithm and Its Variants in Digital Image Processing: A Comprehensive Review," in *Applications of Firefly Algorithm and its Variants*, Dey, N. Ed. Singapore, Springer, 2020, ch. 1, pp. 1-28, doi: https://doi.org/10.1007/978-981-15-0306-1_1.

[49] N. Dey, A. S. Ashour and S. Bhattacharyya, *Applied nature-inspired computing: algorithms and case studies*, Singapore: Springer International Publishing, 2020.

[50] M. B. Imamoglu, M. Ulutas and G. Ulutas, "A new reversible database watermarking approach with firefly optimization algorithm," *Mathematical Problems in Engineering*, vol. 2017, 2017, doi: https://doi.org/10.1155/2017/1387375.

[51] A. Ghardallou, A. Kricha, A. Sakly and A. Mtibaa, "Adaptive block sized reversible watermarking scheme based on integer transform," in *2016 17th IEEE International Conference on Sciences and Techniques of Automatic Control and Computer --Engineering (STA)*, Sousse, Tunisia, 2016, pp. 347-351.

[52] T. Naheed, I. Usman, T. M. Khan, A. H. Dar and M. F. Shafique, "Intelligent reversible watermarking technique in medical images using GA and PSO," *Optik,* vol. 125, no. 11, pp. 2515-2525, 2014, doi: https://doi.org/10.1016/j.ijleo.2013.10.124.

[53] S. Sharma and H. Patil, "Secure data hiding scheme using firefly algorithm with hidden compression," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 23, no. 2, pp. 525-534, 2020, doi: https://doi.org/10.1080/09720529.2020.1729502

[54] M. K. Pandey, G. Parmar, R. Gupta and A. Sikander, "Lossless robust color image watermarking using lifting scheme and GWO," *International Journal of System Assurance Engineering and Management*, vol. 11, no. 2, pp. 320-331, 2020, doi: https://doi.org/10.1007/s13198-019-00859-w.

[55] H. Zarrabi, A. Emami, P. Khadivi, N. Karimi and S.Samavi, "BlessMark: a blind diagnostically-lossless watermarking framework for medical applications based on deep neural networks," *Multimedia Tools and Applications*, vol. 79, no. 31, pp. 22473-22495, 2020, doi: https://doi.org/10.1007/s11042-020-08698-9.

[56] F. Y. Shih, X. Zhong, I. C. Chang and S. Satoh, "An adjustable-purpose image watermarking technique by particle swarm optimization," *Multimedia Tools and Applications*, vol. 77, no. 2, pp. 1623-1642, 2018, doi: https://doi.org/10.1007/s11042-017-4367-9.

[57] M. Arsalan, A. S. Qureshi, A. Khan and M. Rajarajan, "Protection of medical images and patient related information in healthcare: Using an intelligent and reversible watermarking technique," *Applied Soft Computing*, vol. 51, pp. 168-179, 2017, doi: https://doi.org/10.1016/j.asoc.2016.11.044.

[58] K. Balasamy and S. Ramakrishnan, "An intelligent reversible watermarking system for authenticating medical images using wavelet and PSO," *Cluster Computing*, vol. 22, no. 2, pp. 4431-4442, 2019, doi: https://doi.org/10.1007/s10586-018-1991-8.

[59] L. M. Vargas, "Watermarking based on Difference Expansion and Genetic Algorithms," in *Second International Conference on Advances In Computing, Control And Networking - ACCN 2015,* Bangkok, Thailand, 2015, pp. 12-16.

[60] R. R. Kumari, V. V. Kumar and K. R.Naidu, "Optimized DWT Based Digital Image Watermarking and Extraction Using RNN-LSTM", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 2, pp. 150-162, 2021, doi: 10.9781/ijimai.2021.10.006

[61] A. Habib and M. Khalil, "QAM-DWT-SVD Based Watermarking Scheme for Medical Images", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 3, pp. 81-90, 2018, doi: 10.9781/ijimai.2018.01.001

[62] S. Kaur, G. Chaudhary, J. D. Kumar, M. S. Pillai, Y. Gupta, M. Khari, V. García-Díaz and J. Parra Fuente, "Optimizing Fast Fourier Transform (FFT) Image Compression using Intelligent Water Drop (IWD) Algorithm," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 3, no. 7, pp. 48-55, 2022, doi: http://dx.doi.org/10.9781/ijimai.2022.01.004

[63] S. Mirjalili, S. M. Mirjalili and A. Lewis, "Grey wolf optimizer," *Advances in engineering software*, vol. 69, pp. 46-61, 2014, doi: https://doi.org/10.1016/j.advengsoft.2013.12.007.

[64] H. Y. Mahmoud, H. M. Hasanien, A. H. Besheer, and A. Y. Abdelaziz, "Hybrid cuckoo search algorithm and grey wolf optimiser-based optimal control strategy for performance enhancement of HVDC-based offshore wind farms," *IET Generation, Transmission & Distribution*, vol. 14, no. 10, pp. 1902-1911, 2020, doi: https://doi.org/10.1049/iet-gtd.2019.0801

[65] C. Banchhor and N. Srinivasu, "Integrating Cuckoo search-Grey wolf optimization and Correlative Naive Bayes classifier with Map Reduce model for big data classification," *Data & Knowledge Engineering*, vol. 127, pp. 101788, 2020, doi: https://doi.org/10.1016/j.datak.2019.101788.

[66] H. Xu, X. Liu and J. Su, "An improved grey wolf optimizer algorithm integrated with Cuckoo Search,", in *2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Bucharest, Romania, 2017, pp. 490-493.

[67] C. Li and A. C. Bovik, "Three-component weighted structural similarity index," in *Image quality and system performance VI*, vol. 7242, San Jose, California, United States, pp. 252-260, SPIE, 2009.

[68] S. P. Maity and H. K. Maity, "Optimality in distortion control in reversible watermarking using genetic algorithms," *International Journal of Image and Graphics,* vol. 17, no. 03, pp. 1750013, 2017, doi: https://doi.org/10.1142/S0219467817500139.

[69] J. T. Akagi, C. F. Gaona, F. Mendoza, M. P. Saikia and M. Villagra, "Hard and easy instances of L-tromino tilings," *Theoretical Computer Science*, vol. 815, pp. 197-212, 2020, doi: https://doi.org/10.1016/j.tcs.2020.02.025.

[70] X. Yang and X. Jiang, "A hybrid active contour model based on new edge-stop functions for image segmentation," *International Journal of Ambient Computing and Intelligence (IJACI),* vol. 11, no. 1, pp. 87-98, 2020, doi: 10.4018/IJACI.2020010105.

[71] A. E. H. Saad, Z. Dong and M. Karimi, "A comparative study on recently-introduced nature-based global optimization methods in complex mechanical system design," *Algorithms,* vol. 10, no. 4, pp. 120, 2017, doi: https://doi.org/10.3390/a10040120.

[72] A. Jain and V. Bhatnagar, "Concoction of ambient intelligence and big data for better patient ministration services," *International Journal of*

Ambient Computing and Intelligence (IJACI), vol. 8, no. 4, pp. 19-30, 2017, doi: 10.4018/IJACI.2017100102.

[73] P. Kaur, S. Gupta, S. Dhingra, S. Sharma and A. Arora, "Towards content-dependent social media platform preference analysis," *International Journal of Ambient Computing and Intelligence (IJACI)*, vol. 11, no. 2, pp. 30-47, 2020, doi: 10.4018/IJACI.2020040102.

[74] S. Khan, L. Han, H. Lu, K. K. Butt, G. Bachira and N. -U. Khan, "A New Hybrid Image Encryption Algorithm Based on 2D-CA, FSM-DNA Rule Generator, and FSBI," in *IEEE Access*, vol. 7, pp. 81333-81350, 2019, doi: 10.1109/ACCESS.2019.2920383.

[75] A. Alghafis, F. Firdousi, M. Khan, S. I. Batool and M. Amin, "An efficient image encryption scheme based on chaotic and Deoxyribonucleic acid sequencing," *Mathematics and Computers in Simulation*, vol. 177, pp. 441-466, 2020, doi: https://doi.org/10.1016/j.matcom.2020.05.016.

[76] Y. He, Y. Q. Zhang, X. Y. Wang, "A new image encryption algorithm based on two-dimensional spatiotemporal chaotic system," *Neural Computing and Applications*, vol. 32, no. 1, pp. 247-260, 2020, doi: https://doi.org/10.1007/s00521-018-3577-z.

Lakshmi H R

Lakshmi H R completed her B. E. from Dayananda Sagar College of Engineering (affiliated to Visvesvaraya Technological University - VTU) in Electronics & Communication. She did her M. Tech in VLSI Design & Embedded Systems from B N M Institute of Technology (affiliated to VTU). She is currently pursuing her Ph.D. from K. S. Institute of Technology Research Center (affiliated to VTU). She has over 6 years of experience in her capacity as Assistant Professor and Researcher with topics of interest being – Image Processing, Information Security, VLSI, Embedded Systems. She has won the Young Woman Educator & Researcher Award by National Foundation for Entrepreneurship Development (NFED). She has authored many papers and book chapters in reputed journals and conferences. She has several patent publications. Her peer recognition includes her professional memberships & services in refereed organizations, program committees and review boards.

Surekha Borra

Surekha Borra (Senior Member, IEEE) received her B.Tech. from Nagarjuna University, India, in 2003. MTech. and Ph.D. from Jawaharlal Nehru Technological University, Hyderabad, India in 2007 and 2015. She started her academic career as Assistant Professor in 2004 and served in various engineering colleges for 18 years. Currently, she is Professor in the Department of Electronics and Communication Engineering, K. S. Institute of Technology, Bengaluru, India. Dr Borra's research interests include Image and Video Analytics, Information Security and Signal Processing. She has received Woman Achiever's Award from The Institution of Engineers (India) for her prominent research and innovative contribution(s), and several research grants from the Government of Karnataka, India.

Reliability of IBM's Public Quantum Computers

Raquel Pérez-Antón[1], Alberto Corbi[2], José Ignacio López Sánchez[1], Daniel Burgos[2] *

[1] Universidad Internacional de La Rioja (UNIR), Logroño (Spain)
[2] Research Institute for Innovation & Technology in Education (UNIR iTED), Universidad Internacional de La Rioja (UNIR), Logroño (Spain)

* Corresponding author: raquel.perez527@comunidadunir.net (R. Pérez-Antón), alberto.corbi@unir.net (A. Corbi), joseignacio.lopez@ unir.net (J. I. López Sánchez), daniel.burgos@unir.net (D. Burgos).

Received 11 November 2021 | Accepted 17 March 2023 | Published 17 April 2023

Abstract

One of the challenges of the current ecosystem of quantum computers (QC) is the stabilization of the coherence associated with the entanglement of the states of their inner qubits. In this empirical study, we monitor the reliability of IBM's public-access QCs network on a daily basis. Each of these state-of-the-art machines has a totally different qubit association, and this entails that for a given (same) input program, they may output a different set of probabilities for the assembly of results (including both the right and the wrong ones). Although we focus on the computing structure provided by the "Big Blue" company, our survey can be easily transferred to other currently available quantum mainframes. In more detail, we probe these quantum processors with an ad hoc designed computationally demanding quaternary search algorithm. As stated, this quantum program is executed every 24 hours (for nearly 100 days) and its goal is to put to the limit the operational capacity of this novel and genuine type of equipment. Next, we perform a comparative analysis of the obtained results according to the singularities of each computer and over the total number of executions. In addition, we subsequently apply (for 50 days) an improvement filtering to perform noise mitigation on the results obtained proposed by IBM. The Yorktown 5-qubit computer reaches noise filtering of up to 33% in one day, that is, a 90% confidence level is reached in the expected results. From our continuous and long-term tests, we derive that room still exists regarding the improvement of quantum calculators in order to guarantee enough confidence in the returned outcomes.

Keywords

Error Mitigation, IBM Q, Quantum Computers, Quality Control, Quantum Search Algorithms.

DOI: 10.9781/ijimai.2023.04.005

I. Introduction

Quantum computing is a very promising and radically incipient area of knowledge in contrast with the current limitations of classical computing. For instance, classic transistors have a finite physical volume, and we are already approaching the 1 nm limit. Theory predicts that after surpassing this physical dimension, manipulating the flow of the electric current (without the loss of electrons) may be operationally impossible [1]. That is why, additional computational technologies may be needed to expand the ability to solve some (new) complex problems or perform extremely convoluted calculations.

According to the scientific community, quantum computing may be the solution for addressing these issues, including new frontier challenges related to machine learning and artificial intelligence [2]. However, despite all the efforts in this discipline, the challenges are still overwhelming. Perhaps, the biggest one has to do with the stability of the quantum states and the way they are organized in an interdependent way (*entangled*). This characteristic (preservation of entanglement) is known as *coherence* and obviously, critically depends on the number of qubits in a quantum computer (QC), since the greater the number of entangled units in the quantum system, the more sensitive the system to external fluctuations. In turn, for a group of entangled qubits, there is no certainty about the exact state in which each of them is at the individual level. It is what is called *superposition of states*. Coherence in quantum computing can be defined as the conservation of the superposition state of a system over time. In a certain way, we could link this property to a loss of the individuality of each unit (qubit), to behave as a whole (hence the term coherence), and it is a requirement of the system. This property causes the system to be extremely sensitive to interferences from the environment, and as coherence can be destroyed by simple mechanical vibrations, electromagnetic disturbances, sound waves, tiny seismic temblors, or even adverse weather effects. When this takes places, the quantum wave functions collapse as if they were being measured, and the system loses its multi-state nature. This unavoidable process is known as *decoherence*. The maintenance of coherence (or avoidance of decoherence) in this type of hardware is essential for the correct implementation and execution of quantum instructions and the derivation of the expected (and accurate) results. However, in practice it has been shown that quantum decoherence can be minimized, but it is not possible, at least today, to eliminate it completely.

Please cite this article in press as:
R. Pérez-Antón, A. Corbi, J. I. López Sánchez, D. Burgos. Reliability of IBM's Public Quantum Computers, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 155-163, 2025, http://dx.doi.org/10.9781/ijimai.2023.04.005

Therefore, quantum error correction (QEC) is necessary in quantum computing to protect information from errors caused by decoherence and other sources of noise at the quantum level (see [3]–[6]). However, while classic error correction uses the redundancy process to counteract errors (storing the information several times in such a way that if the copies are not the same later, you can choose the option that is generally present) this possibility is no longer feasible in quantum computing according to the no-cloning theorem [7]. Even though this theorem seems to present an obstacle to formulating a theory of QEC, some alternative strategies exist. One of these has to do with the spread of qubits through highly entangled states of several neighboring physical qubits in such a way that a state inversion event could be detected without the need for consulting the exact value of the examined qubit (which would destroy the information). These qubit aggregates (making up a compound logical qubit) are resistant to errors in the final computer. Clearly, this means that if a program requires 10 qubits to run, in practice, it will need 10 logical qubits, which can be translated into hundreds or thousands of the original physical ones. These systems, called *noisy intermediate scale quantum computers* or NISQ, are expected to provide the advantages necessary to meet the required QEC [8], [9].

All errors can be corrected if the imperfections of quantum operations and measurements are below a certain threshold and the correction can be applied repeatedly [10], [11]. However, these error thresholds also depend on the details of the physical system and quantifying them requires careful analysis of both the hardware and software implementation [12].

Although some studies have addressed the quantification of these error baselines for different platforms and QC configurations (see [13]), it is a relatively new area which needs further clarification through experimentation. For this reason, in this work, the reliability of the coherence of IBM's public quantum computers has been examined. The choice of the Big Blue's network of QCs has not been arbitrary. Two factors have influenced this decision. On one hand and for several years, IBM has provided, free of charge, some of its QC infrastructure for research and study. On the other, each piece of equipment is designed differently and with a contrasting qubit number, arrangement, and entangling layout, which determined the possibility of conducting these tests in a variety of configurations for comparative purposes.

In more detail, our experiment consisted of executing for almost 100 days, 1024 times each day, the same quaternary search algorithm (described in Section IV) on 8 IBM public quantum computers. For the sake of completeness, the characteristics of IBM's public quantum processors as well as the environmental conditions in which they are designed to operate are tackled in Section III. Our results are then presented and discussed in Section VI. Finally, some conclusions are drawn in Section IX.

II. Previous Works

Closely related to our work, other very recent research efforts have carried out a verification process of the reliability of IBM quantum processors applying different study methodologies both on the number of qubits and quantum gates. The depth algorithm fragmentation method used in [14] is applied 8192 times to 20 quantum processors in a single day, showing that recomposing fragmentations significantly mitigates noise and decoherence. On the other hand, the work carried out by [15] is based on the study of non-resonant holonomic gates of 3 qubits on the resonant ones. Demonstrating that 3-qubit non-resonant holonomic gates show higher fidelity (80%) compared to resonant gates (70%). In the experiment carried out in [16], its authors focus on a single 5-qubit quantum computer to create during an evaluator scenario three models of state evolution such as inversion recovery, Ramsey and entanglement-deentanglement. They conclude that the framework of steepest entropy-ascent quantum thermodynamics (SEAQT) can be used as a basis for error mitigation schemes. In contrast, the work carried out by [17] on a 20-qubit computer shows that the application of Fourier transforms can be taken as filters that improve the oscillation patterns of the expected data.

III. IBM's Public Quantum Processors

In 2016, IBM deployed and made publicly available the first 5-qubit cloud QC. This was followed by others that were organized into families according to their number of qubits. Each family was named after a bird (Table 1). Thus, we have the 5-qubit processors, which formed the Canary family, the 16-qubit processors such as Albatross, Penguin with 20 qubits, etc. In addition, within each family, the processors are named after a city, so within the Canary family are London, Rome, Vigo, etc. Melbourne is a 16-qubit QC included in the Albatross family. These names are usually given in a personal and loving way by the specific team behind the design and assemble of each computer.

TABLE I. List of Names and Categories According to the Number of Qubits of IBM Computers [18]

Category	Qubits	Processors
Canary	5	Tenerife, Yorktown, Ourense, London, Vigo, Rome, Burlington, Valencia, Santiago
Albatross	16	Melbourne
Penguin	8–16	Austin, Tokyo, Poughkeepsie, Johannesburg, Singapore, Almaden, Boeblingen
Hummingbird	+16	Raleigh

A. Main Characteristics and Components of the IBM Q Equipment

As it is shown in Table II, the characteristics of each quantum computer in which the experiment (detailed in Section 4) has been carried-out, as well as their corresponding numbers of qubits. The type of gates that were used for the design and construction of their circuits is also included. In more detail, u1, u2, and u3 are the three parameters that allow the building of any single qubit gate and have a duration of one unit of time [19]. In addition, the error rate that each door could

TABLE II. Qubit Error Rate and the Characteristics of Each Active Public IBM Q Experience Processor

Name	Qubits	Error Rate Door CNOT	Basic Doors	Single-qubit u2 Error Rate
Melbourne	15	2.384e-2/1.000e+0	id, u1, u2, u3, cx	4.632e-4/3.482e-2
London	5	9.411e-3/1.430e-2	u1, u2, u3, cx, id	3.369e-4/4.624e-4
Burlington	5	9.009e-3/2.075e-2	u1, u2, u3, cx, id	3.568e-4/6.144e-4
Essex	5	8.434e-3/1.406e-2	u1, u2, u3, cx, id	3.929e-47.155e-4
Ourense	5	6.851e-3/2.976e-2	u1, u2, u3, cx, id	2.961e-4/9.845e-4
Vigo	5	7.772e-3/1.470e-2	u1, u2, u3, cx, id	3.673e-4/8.616e-4
Yorktown	5	1.280e-2/2.203e-2	u1, u2, u3, cx, id	6.106e-4/7.950e-4

develop at the time of the measurement is also referenced. This ratio will increase as time passes if the QC is not properly calibrated.

Specifically, IBM performs these calibrations twice a day on each quantum processor and conveniently keeps the users informed so that they can take them into account when eventually launching their programs. Calibration consists of carrying out a series of experiments to obtain precise information about the physical behaviour of each qubit. The values of the parameters that characterize a qubit are different for each qubit within the processor and among different processors, and these can even vary over time. It is possible to identify the qubit's proper frequency by sweeping through a range of frequencies and observing absorption signals. The qubit's frequency is the energy difference between the ground state and the excited state. Aside from calibration, these processors must remain in specific environmental conditions. They also need a temperature close to absolute zero 0 K (−273.144 °C) to better account for the Heisenberg's uncertainty principle [20].

IBM public quantum computer hardware uses the characteristic known as superconductivity. The materials with this property can carry electrical currents without the resistance or loss of energy under specific circumstances. From an architectural point of view, superconductors are wrapped in the form of Josephson joints [21]. These structures (Fig. 1) are formed using two sheets of aluminum, which, under normal environmental circumstances, would behave like classical electrical circuits. However, in the subatomic world, they operate as quantum gates.

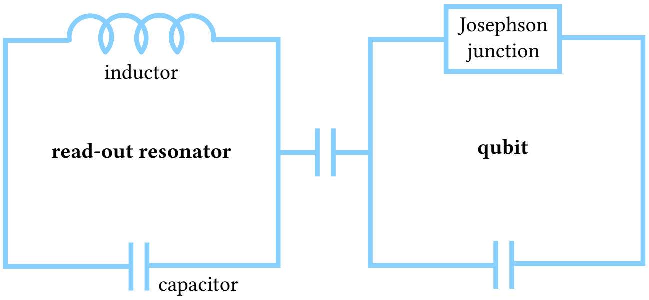


Fig. 1. Superconductivity diagram with Josephson joint (1 qubit).

The transition between the possible states of the qubits is generated by applying a certain level of energy, and because of the tunneling effect, the particle crosses the barrier (with some probability). The state of the qubit can be read by observing the energy of each aluminum sheet, which in turn, causes the decoherence of the system as a whole, but allows us to obtain information about the state in which each constituent unit (qubit) remained after the process.

B. Quantum Computer Connectivity

As stated above, each IBM public quantum processor has a different physical architecture. Nevertheless, the logic of any given quantum algorithm can be applied independently of the subjacent hardware. Even so, it is important to know the internal structure of these computers for several reasons. To begin with, the circuit will use, at most, all the qubits of the processor only once since the algorithms are according to the principles of their construction [22], e.g., principle 3: *long relevant decoherence times much longer than the gate operation time*). Furthermore, the application of a logic gate to several qubits requires, for greater efficiency, that they be physically interconnected in their architecture, because even if the phenomenon of entanglement itself is not conditioned by distance (two qubits could be mutually entangled even at distant points in our universe), our technological capacity to govern that entanglement to our convenience, is critically restricted by its physical separation. Therefore and a priori, the more qubits a processor has and the greater their interconnection, the better the expected results.

As can be seen in Fig. 3, Fig. 2 and Fig. 4, each IBM public QC refers to a different connectivity or type of entanglement depending on the number of qubits and their arrangement. Each circle represents a qubit, and from this possible entanglement lines emerge towards the contiguous qubits. The evolution in connectivity and lattice design is one of IBM's ongoing investigations, as shown in [18]. The debugging of errors in the gates and exposure to crosstalk is linked to the connectivity among the qubits. Therefore, the new processors are built on improvements in previous structural experiences.

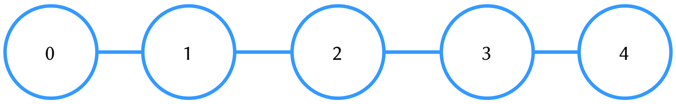


Fig. 2. Qubit arrangement for the Rome quantum computer. The structure of this processor is iterative, linking all the qubits of the processor in an orderly fashion.

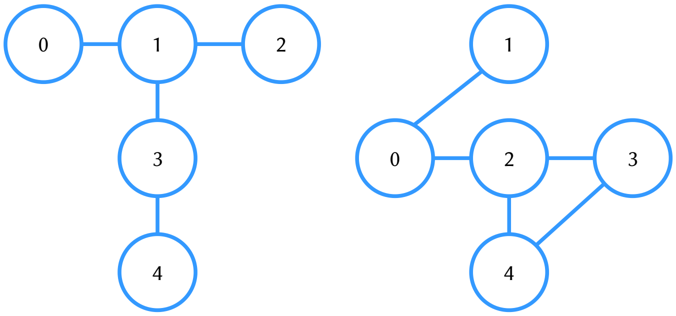


Fig. 4. Qubit arrangement for the London, Burlington, Essex, Ourense (left) and Yorktown (right) quantum computers. These architectures are a composite of Melbourne and Rome since they combine the lattice structure with the iterative one.

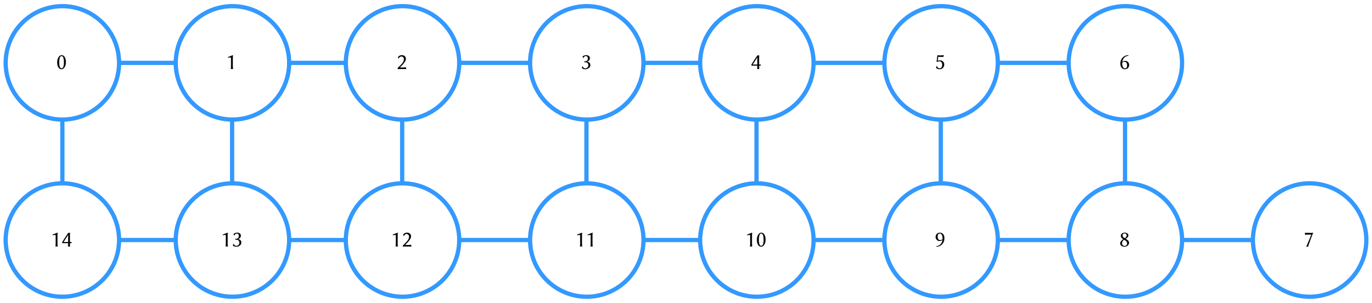


Fig. 3. Qubit arrangement for the Melbourne quantum computer. This architecture is used on processors with more than 16 qubits.

IV. Design of the Quaternary Search Algorithm

In classical computing, a binary tree is a data structure widely used in dynamic memory programming. Each node of the complete tree can have a left and a right child, where its complexity in the search for ordered elements in the best case is as follows:

$$(O(\log_2(N)) \tag{1}$$

where N is the number of nodes in (1). The algorithm presented here uses the data structure of a tree, but in this case, it exploits the intrinsic characteristics of the entanglement of qubits, thus managing a quaternary tree. Each node has four children. The complexity associated with searching in this structure will be considerably reduced in the best case if we transform it into a quaternary tree [23]:

$$O\left(\frac{1}{4\pi}\cdot \ln(N) - 1\right) \tag{2}$$

Besides, our search algorithm forces an iterative entanglement to verify the consistency and stability of the qubits. Therefore, if we take as a reference Grover's basic search algorithm [24], the number of iterations is equal to:

$$\sum U_f = (n_{\text{qubits}}) - 1 \tag{3}$$

where U_f is the so-called oracle (i.e. the unitary operator), and f is a Boolean function. According to (3), each iteration performs the addition of amplitudes until it approaches 1. As we can see, n_{qubits} will need $n-1$ iterations to find the element of the list regardless of whether the first or last element is found. On the contrary, if we search a quaternary tree using the initial state of entanglement of two qubits {00, 01, 10, 11}, as shown in Fig. 5, the number of oracles to be used will be equal to the number of levels in the tree:

$$\sum U_f = \sum L \tag{4}$$

where L is the number of levels in the tree and U_f is the number of oracles in (4). Furthermore, each oracle will return the maximum possible amplitude, so unlike Grover's algorithm we only need to apply the oracle once in each iteration. In our case, we have reduced the application of oracles to two iterations. The result of each oracle is concatenated with the next one until they are finalized in a leaf of the tree. The entanglements of the qubits in each of the oracles are detailed next:

- $L1$ = {00, 01, 10, 11}, where the entanglement is formulated by $q[3]$ and $q[2]$. We generate a vector of states in the 4-D Hilbert space (H^4). The expression $q[i]$ for qubit q and i is the position of the qubit in the circuit.
- $L2$ = {0000, 0001, 0010, 0011, 0100, 0101,...}, where the entanglement will be formulated by $q[3]$, $q[2]$, $q[1]$, and $q[0]$, if and only if the element has not been found in $L1$. We generate a vector of states in the 16-D Hilbert space.

To achieve this, first the Pauli gate X has been applied to $q[2]$, with the aim of changing the initial state. Then the qubits $q[3]$ and $q[2]$have been entangled applying the following Hadamard gates that perform a rotation π around the XZ axis:

$$H|0\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle) \qquad H|1\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle) \tag{5}$$

From here, we can write (5) as the matrix given by (6):

$$H = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix} \tag{6}$$

The state of entanglement of two qubits is determined by the Einstein–Podolsky–Rosen (EPR) pair [25]:

$$|q[2]\,q[3]\rangle = \frac{|00\rangle + |11\rangle}{\sqrt{2}} \tag{7}$$

Therefore, this state $|q[2]\,q[3]\rangle$ described in (7) cannot be decomposed into pure states since no combination of complex coefficients fulfills both descriptions. Therefore, as an alternative to assembling pure states, it is possible to describe mixed states through the matrix or density operator ρ, explained in [26]. We then define the assembly of pure states as the set $\{\rho_i \mid \psi\}$ where ρ_i are all possible states of and thus $\sum\rho_i = 1$. Then the density operator or density matrix is the result of the entanglement of several qubits:

$$\rho = \sum_i \rho_i\, |\psi_i\rangle\langle\psi_i| \tag{8}$$

If we write (8) in matrix form, we have:

$$\rho = \frac{1}{N} I \tag{9}$$

where N is the number of possible states in its measurement, and I is the identity matrix in (9). For example, the density matrix for a single qubit will be $\rho = \frac{1}{2}\left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)$. In our case, with 2 qubits, the generated initial density matrices will be:

$$\{\rho \mid q[2]\,q[3]\} = \frac{1}{4}\begin{pmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{pmatrix} \tag{10}$$

If we write (10) for 4 qubits:

$$\{\rho \mid q[2]\,q[3]\,q[1]\,q[0]\} = \frac{1}{16}\begin{pmatrix}1 & \cdots & 0\\ \vdots & 1 & \vdots\\ 0 & \cdots & 1\end{pmatrix} \tag{11}$$

where the set $\{\rho \mid q[0] \ldots q[n]\}$ defines the state probabilities of the entangled qubits. In our case, we want to find state 01 at level $L1$ and then state 0111 at level $L2$. Hence, we change the sign of the amplitude of the states we are looking for such that:

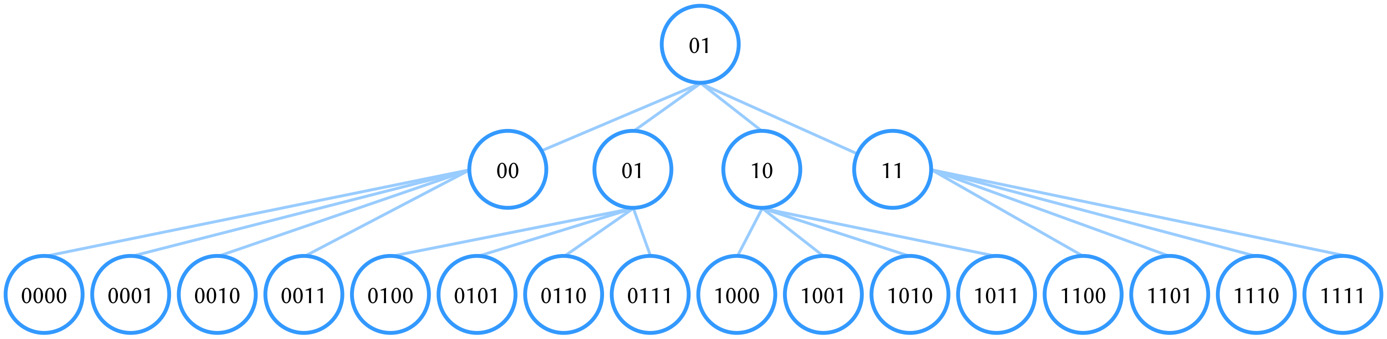


Fig. 5. Graph representation of the complete quantum quaternary tree.

$$
\begin{gathered}
U_{01}|01\rangle = -|01\rangle\,, \\
U_{01}|q[2]\,q[3]\rangle = |-q[2]\,q[3]\rangle\,, \\
U_{0111}|0111\rangle = -|0111\rangle\,, \\
U_{0111}|q[2]\,q[3]\,q[1]\,q[0]\rangle = -|q[2]\,q[3]\,q[1]\,q[0]\rangle \\
\forall \quad q[2]\,q[3]\,q[1]\,q[0] \neq 0111
\end{gathered} \tag{12}
$$

After applying (12), the unit transformation of the oracle U is applied. The oracles that must be applied to the algorithm are scalar according to the number of search qubits. In addition, we will only need one oracle for each level, thus reducing the Grover's algorithm:

$$
L1 = (I-2)|\omega_0\rangle\langle\omega_0|) \qquad L2 = (I-2)|\omega_1\rangle\langle\omega_1|) \tag{13}
$$

The oracles for 2- and 4-qubit entanglement are:

$$
U_{\omega_0} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{14}
$$

$$
U_{\omega_1} = \begin{pmatrix} 0_{0000,0000} & \cdots & 0_{0000,1111} \\ \vdots & 1_{0111,0111} & \vdots \\ 0_{1111,0000} & \cdots & 0_{1111,1111} \end{pmatrix} \tag{15}
$$

where $\omega_0 = 01$ and $\omega_1 = 0111$. The resulting algorithm forces the qubits to iteratively intertwine during execution. In other words, the entanglement result of the first two qubits of the first level $L1$ (given by (14) of the tree) will continue to entangle with the second level $L2$ (given by (15)) and so on (if we expand the tree and, consequently, the number of qubits). Therefore, it should be noted that the qubits are not all initialized entangled: additions are made to the initially entangled source. One of the necessary conditions to generate a quantum circuit is the lack of breaks in the code. In this quaternary search algorithm, searching for an element of the first level $L1$ given by (14) would be impossible. All oracles must be evaluated, and the result obtained is stored in a sheet of the second level $L2$ given by (15).

After that, we apply Grover's star operator:

$$
G_f = (2|\omega_0\rangle\langle\omega_0| - I) \tag{16}
$$

Equation (16) is used to increase the amplitude of the element to be found. Recall that Grover (given by (13)) performs a search on unordered items, but in our case, they are ordered. Therefore, in principle, applying this oracle should ensure a successful outcome for the desired element. The operator defined in Grover's algorithm to increase the amplitude consists of applying the inverse of U_f, in our case $L1^{-1}$ and $L2^{-1}$. Considering the following assertions:

$$
\begin{aligned}
&L1 = (I - 2|\omega_0\rangle\langle\omega_0|) && L2 = (I - 2|\omega_1\rangle\langle\omega_1|) \\
&L1^{-1} = (2|\omega_0\rangle\langle\omega_0| - I) && L2^{-1} = (2|\omega_1\rangle\langle\omega_1| - I) \\
&|S_0\rangle = |q[2]\,q[3]\rangle && |S_1\rangle = |q[2]\,q[3]\,q[1]\rangle
\end{aligned} \tag{17}
$$

where $|S0\rangle$ and $|S1\rangle$ in (17) are auxiliary notations for the entanglement states, then the resulting equation of the algorithm is:

$$
\begin{aligned}
L1|S_0\rangle L1^{-1} + L2|S_1\rangle L2^{-1} &= \frac{1}{4\sqrt{4}}\left((4-4)\sum_{S_0 \neq \omega_0}|S_0\rangle + 8|\omega_0\rangle\right) \\
&+ \frac{1}{16\sqrt{16}}\left((16-4)\sum_{S_1 = \omega_1}|S_1\rangle + 44|\omega_1\rangle\right)
\end{aligned} \tag{18}
$$

A graphical and step-by-step representation of the execution of this quaternary search algorithm (given by (18)) is shown in Fig. 6. The Qiskit framework [27] generates this graphical timeline automatically from the Python code available at GitHub. IBM has brought QC closer to the public by giving access to its processors and providing a series of intuitive tools for conducting experiments. In addition, it announced in 2020 the Quantum Educators program which introduced training in this discipline in the classroom. To complete the teaching material, IBM offers the open-source textbook *Learn Quantum Computing Using Qiskit*. Thanks to the initiatives of the Big Blue, many students are able to train in this discipline, which would otherwise be impossible for them [28]. The code in Listing 1 shows the OpenQASM code behind the graphical representation. Both in Fig. 6 and Listing 1, we see the zero entry of the 4 qubits used defined as $q[0]$, $q[1]$, $q[2]$ and $q[3]$. After that, the following methodology is used on the timeline:

- Steps a and b: We initialize the qubit $q[3]$ to one, applying the Pauli gate X (U3) and then we interlace the states of the qubits $q[2]$ and $q[3]$ by applying the Hadamard operator (U2).
- Step c: We apply the gate CZ (Pauli Z (U1) conditioning factor) where the state we want to find is activated, in our case 01.
- Steps d, e, f, g and h: We combine the Pauli Z (U1) and X (U3) doors to perform the unitary operator and its inverse.
- Step i: We introduce the second-level qubits of the trees $q[0]$ and $q[1$ with one applying the Pauli gate X (U3).
- Steps j, k, l, m and n: We combine the Pauli Z (U1) and X(U3) doors to perform the unitary operator and its inverse.
- Steps o, p, q and r: We measure the output of each qubit.

Listing 1. OpenQASM code of the quaternary search algorithm presented in this research work (equivalent to Fig. 6). Comments (lines beginning with #) signal the start of the carried out steps.

```
include ”qelib1.inc”;
qreg q[15];
creg c0[4];
barrier q[0], q[1];
barrier q[0], q[1];
barrier q[0], q[1];
barrier q[0], q[1];
barrier q[0], q[1];
barrier q[0], q[1];
barrier q[0], q[1];
barrier q[0], q[1];
#-----------------------
a u3(3.14, 3.14, 3.14)
q[0]; u3(1.57, 3.14, 0)
q[1]; cx q[0], q[1];
#-----------------------
b barrier q[0];
u1(3.14) q[0];
u2(0, 3.14) q[1];
u2(0, 3.14) q[2];
u3(3.14, 0, 3.14) q[3];
#-----------------------
c cx q[2], q[3];
u3(1.57, 6.28, 3.14)
q[2]; u2(0, 3.14) q[3];
cx q[2], q[3];
u2(0, 3.14) q[2];
barrier q[2], q[3];
u1(3.14) q[2];
u2(0, 6.28) q[3];
cx q[2], q[3];
u2(0, 3.14) q[2];
#-----------------------
d cx q[1], q[2];
u2(0, 6.28) q[1];
cx q[0], q[1];
barrier q[0];
u2(0, 3.14) q[1];
u2(3.14, 3.14) q[2];
u2(0, 3.14) q[3];
barrier q[3];
u2(0, 6.28) q[3];
cx q[2], q[3];
u2(0, 3.14) q[2];
cx q[1], q[2];
u2(0, 3.14) q[2];
u2(0, 3.14) q[3];
barrier q[3];
#----------------- o
p q r measure q[0] ->
c0[0]; measure q[1] ->
c0[1]; measure q[2] ->
c0[2]; measure q[3] ->
c0[3];
```

V. Data Collection Phase

As explained in Section I, the algorithm described above has been run on a daily basis for almost 100 days. On each execution, each of the IBM QCs (introduced in Section III.B) performed the quaternary search 1024 times. Some data is missing because of occasional maintenance/offline periods. In addition, we have considered the calibration timetable of each piece of hardware by launching our probing code both before and after this housekeeping phase. However, as we can see in Fig. 7, no improvement has been observed in the search for the desired data after the daily calibration.

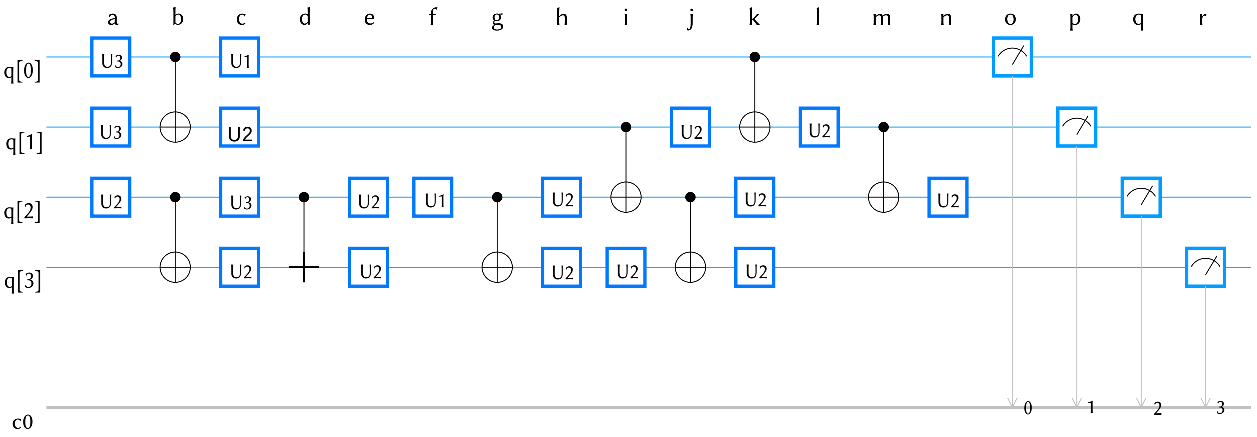


Fig. 6. Graphical representation (IBM OpenQASM 2.0 specification [27]) of the quaternary search algorithm used to probe the stability of quantum hardware. The nomenclature used by OpenQASM 2.0 indicates each turn on the qubit made by the unitary matrix or gate applied in the U form U(theta, phi, lam); where U2 corresponds to the Hadamard gate, U3 to Pauli X gate and U1 to Pauli Z gate.

Each run job produced, among other outcomes, a daily histogram (like the examples shown in Fig. 7 and in Fig. 8) with the probabilities of the sought result. The data of all the probabilities of the elements have been recorded to be later analyzed.

Through a graphical interface or manual code insertion with a simple Jupyter notebook written in the high-level Python language, real qubits have been used, the algorithm has been run online, and experiments have been carried out on these processors. Next, we detail the daily results of the behavior of each remote quantum computer, in addition to making a generic evaluation of all of them.

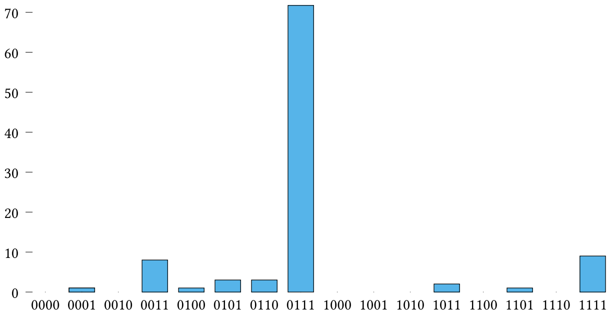


Fig. 7. Histogram of results obtained in the Yorktown QC on June 10, 2020 (after calibration). The 0111 value is the correct one.

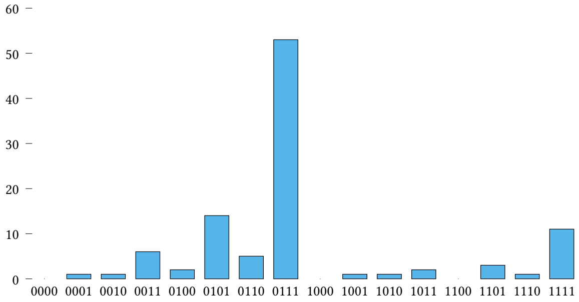


Fig. 8. Histogram of results obtained in the Burlington QC on May 29, 2020 (after calibration). The 0111 value is the correct one.

As seen in Fig. 2, Fig. 3 and Fig. 4, each computer registers a certain initial interlacing architecture. However, the execution process is supposed to generate all the necessary interleaves for any given execution. Table III shows the relationship between the original entanglements for each computer and the proposed quaternary search algorithm.

TABLE III. Entanglement Relationship Between the Intrinsic Architecture of Each Quantum Computer and the Entanglement Forced by the Search Algorithm (Initial Processor Interleaving Architecture). However, There Are Possible Partial Entanglements Indicated in the Third Column

	$q[3]$ & $q[2]$	$q[0]$ & $q[1]$	($q[3]$ & $q[2]$) & ($q[3]$ & $q[2]$)
Melbourne	Yes	Yes	Yes
Rome	Yes	Yes	Yes
London	No	Yes	(q[2])-(q[0]-q[1])
Burlington	No	Yes	(q[2])-(q[0]-q[1])
Essex	No	Yes	(q[2])-(q[0]-q[1])
Ourense	No	Yes	(q[2])-(q[0]-q[1])
Vigo	No	Yes	(q[2])-(q[0]-q[1])
Yorktown	Yes	Yes	Yes

VI. Results

Next, we detail the behavior of each piece of remote hardware when running the quantum program described in Section IV. It is possible to establish two main categories: those who show a low variability and those who exhibit erratic performance over time.

A. Quantum Computers With a Stable Trend Over Time

As shown in Fig. 9, the Yorktown computer presents the best trend to reach the probability (66%) of the correct result (0111). In addition, the difference between all the obtained values obtained does not exceed 5%. On the other hand, Fig. 10 shows the results obtained through the daily observation of the execution in this remote computer.

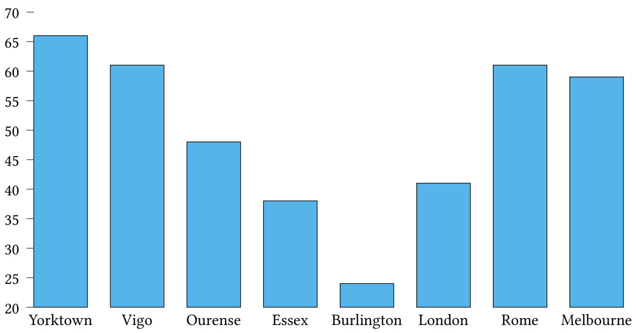


Fig. 9. Average probability of the desired result (0111) through the ~90 days of running the algorithm on several remote QC.

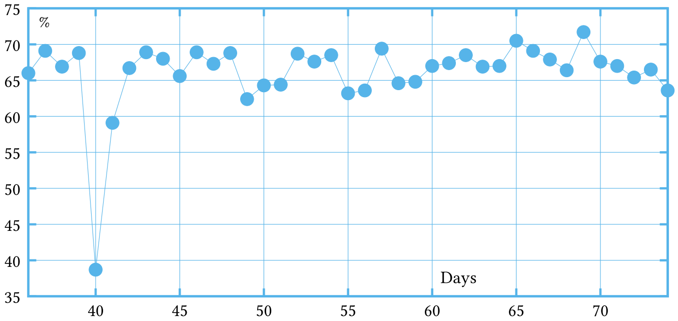


Fig. 10. Results of the Yorktown IBM equipment. Each data point refers to the probability of the desired result (0111) through the days of running the algorithm on this remote quantum computer.

B. Unified Evaluation of All the Quantum Computers

In Fig. 9, it is shown a global (averaged over time) view of the performance of the 8 QCs. As we can see, none of them exceed the 65% average probability. Furthermore, the computer with the highest number of qubits (Melbourne) does not offer the greatest performance, which seems to indicate the importance of decoherence of entangled qubits during the process.

VII. Mitigation of Crosstalk

The term Noisy Intermediate Scale Quantum (NISQ) refers to prototype systems with 5-20 qubits that are now available for wide public use [9], as is the case of those that have been used in this work. In NISQ systems, a major source of noise such as diphony corrupts quantum states when multiple gates or instructions are executed simultaneously [29]. Noise reduction through crosstalk mitigation in IBM quantum computers is done physically on the hardware through daily calibrations. However, this calibration is impossible for us to carry out (since we do not have privileged access to the hardware) and therefore, in this work other mitigation methods are explored through the application of software. One of the basic software tools proposed by IBM for noise mitigation in its quantum computers is the application of filters through noise matrices [30]. IBM's proposal to reduce noise is carried out through Qiskit's open `CompleteMeasFitter` class and consists of applying software filters on the initial probabilistic results. These filters are based on the creation of a noise matrix, which houses the deviations from the basic states. Therefore, any other state in superposition will be helped by a weighting in these deviations. Detailed development of this methodology can be found in the open IBM Q Experience documentation as *Measurement Error Mitigation*. The noise mitigation software has been applied to the same algorithm described in Section IV for 50 days, which was carried out, as an example, on the Santiago, Bogotá and Yorktown instruments. The reason for running this process over several days is based on the variability of daily calibrations that IBM performs at its facilities. That is, the noise matrix applied to the algorithm is different in each execution, as are the probabilistic results of the quaternary search.

The hypothesis test statistic applied for this case is Wilcoxon [31], since the test variables are adjusted to its methodology. The hypothesis proposed suggests a significant improvement in the initial results by applying noise mitigation. As can be seen in the box diagram in Fig. 11, the combined improvement exceeds 85% of the expected value. Therefore, we can conjecture that the application of noise mitigation in measurements proposed by IBM gives the expected results.

VIII. Discussion

According to the results obtained in Section VI (when compared against the expected ones), we can derive that a greater number of qubits does not guarantee a better response of a quantum computer. This is due to different reasons. In the first place, the entanglement configuration between different qubits is a determining issue. In our case, and as described in Section IV, our algorithm forces the qubits $q[3] - q[2]$ and $q[0] - q[1]$ to be entangled. However, only the initial interleaving architecture of the Melbourne, Rome and Yorktown computers meet this requirement, as shown in Table III and Fig. 9 (with a 60%, 62% and 66% probability, respectively). Regarding the computers London, Burlington, Essex and Ourense, where their initial interleaving configuration is the same, we verify that the results are the worst in the test, as shown in Fig. 9. This circumstance reaffirms our hypothesis about the interleaving architecture required by the tested algorithm and its direct relationship with the distilled results. In addition, we find out that not only the entanglement arrangement influences the initial structure, but also the number of qubits significantly influences it. For instance, the Melbourne computer, despite being the one with the highest number of qubits and having the required structure, it only reached a 60% probability for the expected result (also shown in Fig. 9). However, as can be seen in Fig. 7, after a hardware calibration carried out internally by IBM, the expected result improves significantly.

Finally, it has been possible to slightly balance the decoherence issue (15% improvement) by applying noise mitigation software to the results obtained (as shown in Section VII). From Fig. 11, we consider it to be a favorable result for the application of quantum computing in problem solving areas.

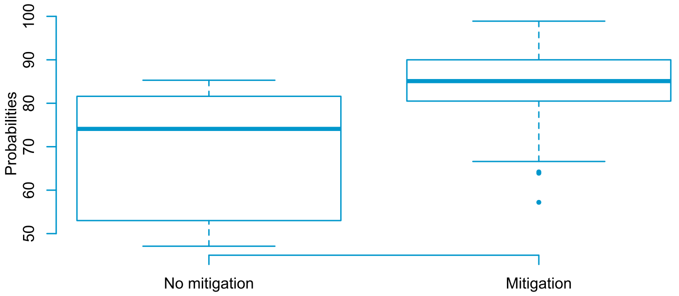


Fig. 11. Result of the application of the Wilcoxon statistic on the noise mitigation values in the quantum computers of IBM Santiago, Bogotá and Yorktown.

IX. Conclusions

In this work, the reliability in time of a specific series of public access quantum processors has been studied through the repeated and transversal execution of the same state-of-the-art quantum algorithm. In addition, a quantum decoherence filtering proposed by IBM has been applied with a significant improvement in the results. The objective was to empirically demonstrate their current suitability for executing and a consuming resource and a computational hungry quantum program. The results obtained provide information on the probability of the correct sequences (in our case known), having shown that the results are highly dependent on the equipment in which they are executed, in turn closely related to the initial configuration of the qubits, their sequence and level of interrelation. Although for a simple task like this, it might seem if we compare with a classical scheme, that the results are not robust enough, extrapolated to tasks beyond the limits of a classical scheme, they reinforce the idea of the great potential that this technology has, even with the need to gain homogeneity over time to guarantee an adequate level of reliability in relevant decisions, such as the business, health, or academic world. On the other hand, its suitability for research, education and the study and advancement of this technology itself has been amply demonstrated (there are currently 380,000 registered users, 1.4 trillion circuits executed and 1400 research articles published).

Future research may, to begin with, expand the number of daily executions of the algorithm. Furthermore, it could also modify the algorithm and assess the level of acceptance of qubit entanglement on the results. It would also be interesting to analyze the level of error of the quantum gates in greater detail. Finally, the fact of not knowing the calibration schedule a priori has also been a limitation.

Acknowledgment

This research is partially funded by the Universidad Internacional de la Rioja (UNIR) through the Research Institute for Innovation & Technology in Education (UNIR iTED). The authors would also like to express their gratitude towards IBM for allowing public access to their hardware and for fostering the free use of their remarkable equipment.

References

[1] J. Gao, M. A. Thompson, *et al.*, *Combined quantum mechanical and molecular mechanical methods*, vol. 712. ACS Publications, 1998.

[2] M. Schuld, I. Sinayskiy, F. Petruccione, "An introduction to quantum machine learning," *Contemporary Physics*, vol. 56, no. 2, pp. 172–185, 2015.

[3] A. Furusawa, J. L. Sørensen, S. L. Braunstein, C. A. Fuchs, H. J. Kimble, E. S. Polzik, "Unconditional quantum teleportation," *Science*, vol. 282, no. 5389, pp. 706–709, 1998.

[4] M. Barrett, J. Chiaverini, T. Schaetz, J. Britton, W. Itano, J. Jost, E. Knill, C. Langer, D. Leibfried, R. Ozeri, *et al.*, "Deterministic quantum teleportation of atomic qubits," *Nature*, vol. 429, no. 6993, pp. 737–739, 2004.

[5] T. Aoki, G. Takahashi, T. Kajiya, J.-i. Yoshikawa, S. L. Braunstein, P. Van Loock, A. Furusawa, "Quantum error correction beyond qubits," *Nature Physics*, vol. 5, no. 8, pp. 541–546, 2009.

[6] M. A. Thornton, "Introduction to quantum computation reliability," in *2020 IEEE International Test Conference (ITC)*, 2020, pp. 1–10, IEEE.

[7] W. K. Wootters, W. H. Zurek, "A single quantum cannot be cloned," *Nature*, vol. 299, no. 5886, pp. 802–803, 1982.

[8] S. S. Tannu, M. K. Qureshi, "Not all qubits are created equal: a case for variability-aware policies for nisq- era quantum computers," in *Proceedings of the Twenty- Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2019, pp. 987–999.

[9] J. Preskill, "Quantum computing in the nisq era and beyond," *Quantum*, vol. 2, p. 79, 2018.

[10] J. Cramer, N. Kalb, M. A. Rol, B. Hensen, M. S. Blok, M. Markham, D. J. Twitchen, R. Hanson, T. H. Taminiau, "Repeated quantum error correction on a continuously encoded qubit by real-time feedback," *Nature communications*, vol. 7, no. 1, pp. 1–7, 2016.

[11] M. Otten, S. K. Gray, "Recovering noise-free quantum observables," *Physical Review A*, vol. 99, no. 1, 2019.

[12] P. Schindler, J. T. Barreiro, T. Monz, V. Nebendahl, D. Nigg, M. Chwalla, M. Hennrich, R. Blatt, "Experimental repetitive quantum error correction," *Science*, vol. 332, no. 6033, pp. 1059–1061, 2011.

[13] N. M. Linke, D. Maslov, M. Roetteler, S. Debnath, C. Figgatt, K. A. Landsman, K. Wright, C. Monroe, "Experimental comparison of two quantum computing architectures," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3305–3310, 2017.

[14] T. Ayral, F.-M. Le Régent, Z. Saleem, Y. Alexeev, M. Suchara, "Quantum divide and compute: Hardware demonstrations and noisy simulations," in *2020 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2020, pp. 138–140, IEEE.

[15] S. Bhattacharyya, S. Bhattacharyya, "Holonomic control of a three-qubits system in an NV center using a near-term quantum computer," *arXiv preprint arXiv:2202.08061*, 2022.

[16] J. Montanez-Barrera, M. R. von Spakovsky, C. Damian- Ascencio, S. Cano-Andrade, "Decoherence predictions in a superconductive quantum device using the steepest- entropy-ascent quantum thermodynamics framework," *arXiv preprint arXiv:2203.08329*, 2022.

[17] A. Francis, J. Freericks, A. Kemper, "Quantum computation of magnon spectra," *Physical Review B*, vol. 101, no. 1, p. 014411, 2020.

[18] J. Chow, J. Gambetta, "Quantum takes flight: Moving from laboratory demonstrations to building systems," *IBM Research Blog*, 2020.

[19] IBM, "IBM Quantum Experience," 2016.

[20] W. Heisenberg, "Schwankungserscheinungen und quantenmechanik," *Zeitschrift für Physik*, vol. 40, no. 7, pp. 501–506, 1927.

[21] B. D. Josephson, "Supercurrents through barriers," *Advances in Physics*, vol. 14, no. 56, pp. 419–451, 1965.

[22] D. P. DiVincenzo, "The physical implementation of quantum computation," *Fortschritte der Physik: Progress of Physics*, vol. 48, no. 9-11, pp. 771–783, 2000.

[23] P. Høyer, J. Neerbek, Y. Shi, "Quantum complexities of ordered searching, sorting, and element distinctness," *Algorithmica*, vol. 34, no. 4, pp. 429–448, 2002.

[24] L. K. Grover, "A fast quantum mechanical algorithm for database search," in *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, 1996, pp. 212–219.

[25] A. Einstein, B. Podolsky, N. Rosen, "Can quantum- mechanical description of physical reality be considered complete?," *Physical review*, vol. 47, no. 10, p. 777, 1935.

[26] J. P. Hecht, *Fundamentos de Computación Cuántica orientados a la criptología teórica*. 2012.

[27] D. C. McKay, T. Alexander, L. Bello, M. J. Biercuk, L. Bishop, J. Chen, J. M. Chow, A. D. Córcoles, D. Egger, S. Filipp, *et al.*, "Qiskit backend specifications for openqasm and openpulse experiments," *arXiv preprint arXiv:1809.03452*, 2018.

[28] M. Tilves, "IBM lleva la computación cuántica a las aulas," *Silicon.es*, 2020.

[29] P. Murali, D. C. McKay, M. Martonosi, A. Javadi-Abhari, "Software mitigation of crosstalk on noisy intermediate- scale quantum computers," in *Proceedings of the Twenty- Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2020, pp. 1001–1016.

[30] IBM, "Measurement error mitigation," 2020.

[31] J. J. Litchfield, F. Wilcoxon, "A simplified method of evaluating dose-effect experiments," *Journal of pharmacology and experimental therapeutics*, vol. 96, no. 2, pp. 99–113, 1949.

Raquel Pérez-Antón

Raquel Pérez-Antón is a PhD Candidate in Computer Science at the Universidad Internacional de La Rioja (UNIR) where she also graduated in Computer Engineering. She also holds a Bachelor's degree in Technical Engineer in Management Computer Science from the Universidad de Alicante (UA), a Master's degree in secondary teaching staff specializing in Mathematics and Computer Science from the Universidad Internacional de Valencia (VIU). She currently works as a secondary school teacher in the Higher Degree in Multiplatform Applications Development (DAM) teaching the Programming, Services and Processes modules, and in the Higher Degree in Web Application Development (DAW) as director of end-of-cycle projects.

Alberto Corbi

Alberto Corbi obtained his PhD in Physics at the Universidad de Valencia (UV) and the Institute for Corpuscular Physics (Spanish Council for Scientific Research). He also works as a senior researcher at the Research Institute for Innovation & Technology in Education (UNIR iTED) and as a professor at the Engineering School, which are both part of the Universidad Internacional de La Rioja (UNIR). He is currently involved in a variety of research fields: eLearning standards, Medical Physics, Radiological Protection, Science Education, monitoring of physical activities, social implications of technology and eHealth advancement (with an accent on Alzheimer's disease and clinical standards). He has published over 20 research papers on all the aforementioned subjects, and he is a frequent speaker and knowledge disseminator at radio stations, podcast shows, scientific workshops, general press, academic settings, and outreach events.

Daniel Burgos

Daniel Burgos is the Vice-rector for International Research, director of the UNESCO Chair on eLearning and of the ICDE Chair on Open Educational Resources, at Universidad Internacional de La Rioja (UNIR). He is also director of the Research Institute for Innovation & Technology in Education (UNIR iTED). His work is focused on Adaptive, Personalised and Informal eLearning, Learning Analytics, eGames, and eLearning Specifications. He has published over 150 scientific papers, 20 books and 15 special issues on indexed journals. He has developed +55 European and Worldwide R&D projects. He holds degrees in Communication (PhD), Computer Science (Dr. Ing), Education (PhD), Anthropology (PhD), Business Administration (DBA), and Artificial Intelligence & Machine Learning (postgraduate, at MIT).

Jose Ignacio López Sánchez

Jose Ignacio López Sánchez obtained his PhD in Chemistry at the Universidad de Murcia (UM), while working for the chemical industry as a Torres Quevedo researcher. Previously, he was awarded with a three-year research grant from the Regional Agency for Science and Innovation from Murcia (Séneca) and had participated in several university-industry research projects. He has held various management positions in the industry as R&D and laboratory director and has participated as co-investigator and PI in regional, national and European projects. As an associate professor at the Engineering School (ESIT), part of the Universidad Internacional de La Rioja (UNIR), he teaches in environment and prevention of occupational hazards. He has published 17 scientific papers in the areas of chemistry and cognitive performance and co-invented national and international patents.

A Multi-Session Evaluation of a Haptic Device in Normal and Critical Conditions: a Mars Analog Mission

Julie Manon[1,2,3,7*], Jean Vanderdonckt[4,5], Michael Saint-Guillain[4], Vladimir Pletser[6], Cyril Wain[7], Jean Jacobs[7], Audrey Comein[7], Sirga Drouet[7], Julien Meert[7], Ignacio Sanchez Casla[7], Olivier Cartiaux[8], Olivier Cornu[1,3]

[1] Neuromusculoskeletal Lab (NMSK), Université Catholique de Louvain (UCLouvain), Brussels/Louvain-la-Neuve (Belgium)
[2] Anatomy and Morphology Lab (MORF), Université Catholique de Louvain (UCLouvain), Brussels/Louvain-la-Neuve (Belgium)
[3] Cliniques universitaires Saint-Luc, Orthopedic Surgery Department, Université Catholique de Louvain (UCLouvain), Brussels/Louvain-la-Neuve (Belgium)
[4] Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), Université Catholique de Louvain (UCLouvain), Brussels/Louvain-la-Neuve (Belgium)
[5] Louvain Research Institute in Management and Organizations (LouRIM), Université Catholique de Louvain (UCLouvain), Brussels/Louvain-la-Neuve (Belgium)
[6] European Space Agency (ret.), Blue Abyss (United Kingdom)
[7] Crew 227 – Mission Analog Research Simulation (M.A.R.S. UCLouvain) – Mars, Desert Research Station (MDRS) Simulation (27 March to 10 April 2022), UT (USA)
[8] Department of Health Engineering, ECAMBrussels Engineering School, Haute Ecole "ICHEC-ECAM-ISFSC", Brussels (Belgium)

* Corresponding author: julie.manon@uclouvain.be

Received 26 May 2023 | Accepted 7 February 2025 | Published 11 April 2025

Abstract

While visual interaction is typically evaluated as an instantaneous, one-shot activity that considers only a snapshot of factors, haptic interaction is more challenging to evaluate as it involves a continuous touch process evolving over time. To better understand how to evaluate haptic interaction, this paper performs a multi-session evaluation of a haptic device to be used by astronauts in future lunar and Mars missions, based on eight factors. Three groups of two members ($n = 6$) applied, either as operator or assistant, a newly developed external fixator (EZExFix) to fix a fracture of the tibial shaft. Astronauts had different levels of expertise, i.e., in anatomy, mechanical engineering, and without, and participated in eight timed runs. Among these eight matches, four sessions were conducted with different time frames and compared to a stress test, a reproduction of the experiment in very stressful conditions, and a session simulating critical conditions in an extra-vehicular activity.

Keywords

Context of Use, Haptic Device, Multi-Session, User Experience, UEQ+.

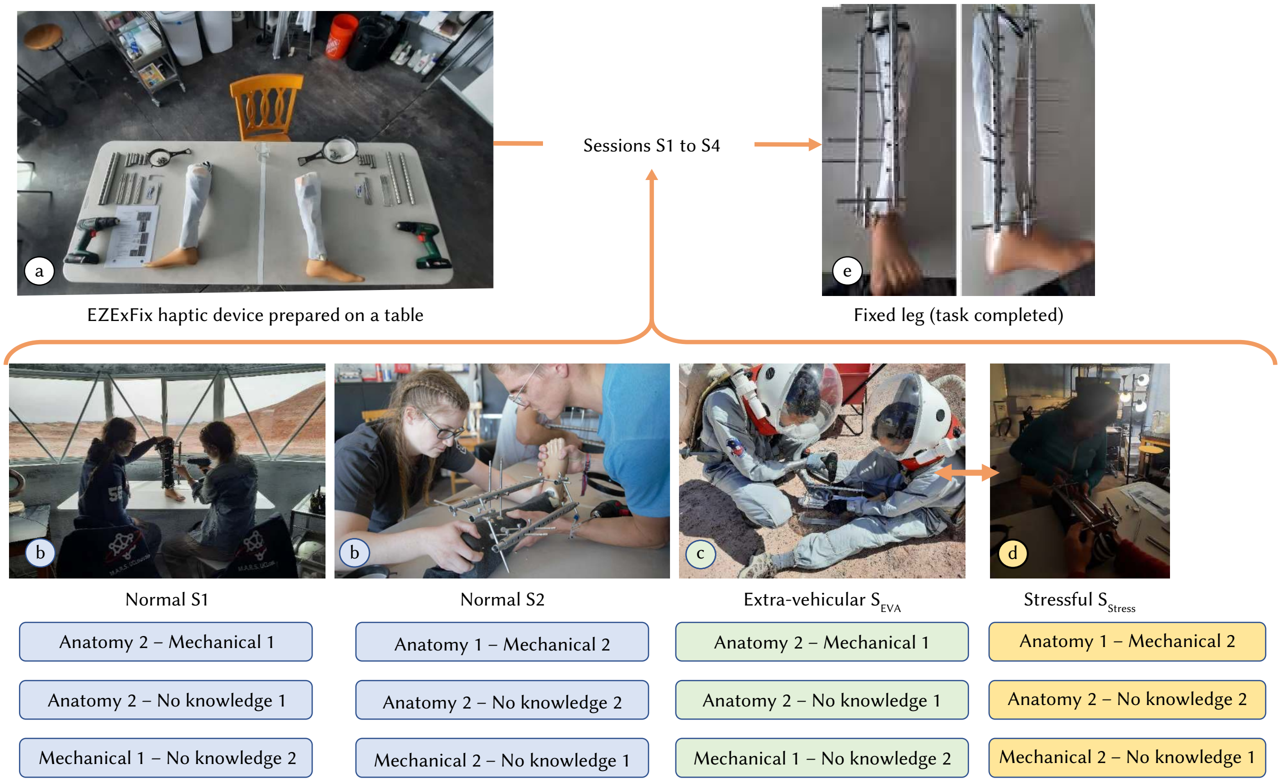
DOI: 10.9781/ijimai.2025.04.001

I. Introduction

Haptic interaction typically promotes the sense of touch as an alternate modality to visual interaction [1] when the visual channel can be occupied, overwhelmed, or simply constrained by other factors, such as in critical conditions. While the visual channel is instantaneous, as for immediate feedback, haptic interaction involves tactile sensations which are part of our somatosensory system, a system that is rather continuous and not as instantaneous as the visual channel. Using a haptic device requires physical manipulation in real time that necessarily involves collision detection and effort to compensate for it. Learning haptic interaction is a continuous process with variations, as for gestural [2] and vocal interaction [3]. For these rea-sons, evaluating the haptic interaction that people can have with a physical device is not just a one-shot action but should be continuously examined over time to capture how people progressively acquire, manipulate, and react to such a haptic device, or, in other words, its evolution over time.

Although there are several methods to quantitatively evaluate a haptic device, they are mostly device-dependent or metric-dependent, which makes them challenging to transpose to another context of use [4]. In contrast, qualitative methods are device-independent but are usually self-reported questionnaires that are limited in scope or not very specific. To better understand how to evaluate a haptic device in

Please cite this article as: J. Manon, J. Vanderdonckt, M. Saint-Guillain, V. Pletser, C. Wain, J. Jacobs, A. Comein, S. Drouet, J. Meert, I. Sanchez Casla, O. Cartiaux, O. Cornu. A Multi-Session Evaluation of a Haptic Device in Normal and Critical Conditions: a Mars Analog Mission, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 164-174, 2025, http://dx.doi.org/10.9781/ijimai.2025.04.001


a
EZExFix haptic device prepared on a table
Sessions S1 to S4
e
Fixed leg (task completed)
b
b
c
d

Normal S1	Normal S2	Extra-vehicular S_{EVA}	Stressful S_{Stress}
Anatomy 2 – Mechanical 1	Anatomy 1 – Mechanical 2	Anatomy 2 – Mechanical 1	Anatomy 1 – Mechanical 2
Anatomy 2 – No knowledge 1	Anatomy 2 – No knowledge 2	Anatomy 2 – No knowledge 1	Anatomy 2 – No knowledge 2
Mechanical 1 – No knowledge 2	Mechanical 2 – No knowledge 1	Mechanical 1 – No knowledge 2	Mechanical 2 – No knowledge 1

Fig. 1. Overview of the experiment: (a) initial preparation of material needed for 2 simultaneous surgeries, (b) sessions covering various conditions, including (c) extra-vehicular activity (EVA) and (d) under stress, (e) final results. Sessions S1 to S4were organized in 12 timed runs (bottom). Six analog astronautswere divided into 3 groups based on their educational backgrounds: with knowledge in anatomy ("Anatomy"), with knowledge in mechanical engineering ("Mechanical"), and without any knowledge in anatomy and engineering ("No knowledge"). Each person was identified by 1 or 2.

this context, we wanted a device that supports haptic assembly [5], as it incorporates a complete haptic effect for mounting and dismounting operations. The task of haptic assembly consists of combining the mechanical joints of a device while focusing on the guidance of objects and the activation signals of the kinematic constraints posed by the device [6].

For this purpose, we selected the EZExFix, a low-cost, fast, and easy-to-use external fixator to handle tibial shaft fractures, which are among the most common open or closed long bone fractures [7][8][9].

The EZExFix consists of a metallic device made up of pins that are inserted into the fractured bone, connecting rods outside the leg (Fig. 1-a and 1-e). The background of its creation, purpose, and validation has been previously published [10]. To compare the operation of such a device in normal and critical conditions, we applied this newly developed EZExFix in realistic operational conditions on Mars during a two-week simulation mission at the Mars Desert Research Station (MDRS, Utah, USA) [11][12].

II. Background

The evaluation of haptic interfaces has been the subject of a great deal of work [13][14] in the context of a range of haptic applications or a particular interactive application with haptic use, mainly in games, virtual reality [4], and machines [6]. For example, Hamam and Saddik [15] pro-posed a mathematical model to evaluate the quality of experience of haptic-based applications, which has been validated through a user study, showing that a Principal Component Analysis performs slightly better than other approaches. Höver et al. [16] presented a user-based evaluation of data-driven haptic rendering, emphasizing the importance of dynamic material effects for achieving realistic haptic feedback. While these studies evaluate the haptic modality in isolation, they acknowledge the need for evaluating both graphical and haptic elements. After reviewing a series of physical and psychophysical metrics used for evaluating a haptic interface [17], Samur derived a psychophysical method for evaluating a force-feedback device [14], which includes guidelines for characterizing such a device along the new dimensions. The specific functions of vibration [13] and sensitivity and friction [18] have been also addressed for a haptic device.

In sum, existing methods focus mainly on the haptic modality, either in general using a model or in particular for a certain type of application in an activity domain. They do not put into perspective evaluation along several dimensions of usability or user experience in a uniform way. For these reasons, we chose a method that evaluates different dimensions in the same way to compare them with each other and across different sessions.

III. Multi-Session Evaluation

A. Participants and Sessions

Three groups of two analog astronauts ($n = 6$) were recruited from the crew 227, who participated in the Tharsis 2022 mission at the MDRS, depending on their level of expertise, established based on their respective degrees or studies: with knowledge in anatomy ("Anatomy"), with knowledge in mechanical engineering ("Mechanical"), and without any knowledge in anatomy and engineering ("No knowledge"). On the first day of the mission, these three groups first attended a short theoretical course on the indications, anatomical landmarks, and steps of EZExFix setup for 1 hour followed by a practical demonstration. Then, they had to perform the task one after another, sometimes playing the role of *operator*, who put the EZExFix on the broken leg (called for example "Anatomy 1" or

"Anatomy 2" depending on the person in the "Anatomy" group - Fig. 1, bottom), sometimes in the role of assistant, who helps to maintain the fracture reduction. Each astronaut met each other in timed runs during which they had to set up the EZExFix to repair an artificial tibial shaft fracture (Fig. 1), in the most efficient way and in the least possible time. Therefore, the number of timed runs consisted of twelve blocks that covered both operators and assistant pairs. Within these blocks, each person was given four times the role of the operator and evaluated on each trial achievement, totaling 24 trials ($N = 24$).

The different groups can also be compared in terms of skills to assess the need to have basics in anatomy or mechanics. Since a fracture could occur in space and therefore be stressful, different conditions were evaluated. The trials were scheduled at the MDRS [19] in good conditions with all the instrumentation prepared on a table (in blue in Fig. 1), during an *extravehicular activity* (EVA) with space suits (in green) or at an unexpected moment, such as at night or dinner, with nothing prepared (in yellow), both considered stressful conditions. The extensive and detailed information has previously been covered [20][21].

B. Design, Measures, and Protocol

Demographic information was collected from the participants before the beginning of the mission. Parti-cipants were instructed to complete a UEQ+ questionnaire (User Experience Questionnaire) [22], a modular extension of the UEQ evaluation method in which eight scales were selected, *i.e.*, Attractiveness, Efficiency, Perspicuity, Trust, Adaptability, Usefulness, Intuitive use, and Haptics, among the 20 scales available to evaluate the user experience of participants interacting with the haptic device. We chose these eight scales for the following reasons: the original UEQ [23] includes Attractiveness as the topmost scale covering (Fig. 2) pragmatic and hedonic qualities, in particular with Perspicuity, Efficiency; we did not consider Dependability, Stimulation and Novelty from the original UEQ as we preferred to focus on more relevant scales and avoid fatigue; we selected the other scales to preserve a balance between these qualities and incorporated explicitly Adaptability to investigate how different users could accommodate the EZExFix, Trust [24] to assess the confidence of the participants, and Haptics [25] because of the haptic nature of the device. These scales typically refer to a question, which is specifically tailored for our experiment:

- Attractiveness: do users like the EZExFix?;
- Efficiency: can users apply the EZExFix without unnecessary effort?;
- Perspicuity: is it easy to get familiar with the EZExFix?;
- Trust: are the users not harmed by the EZExFix?;
- Adaptability: can the EZExFix be adapted to various working styles?;
- Usefulness: does using the EZExFix bring benefits?;
- Intuitive use: can the EZExFix be used immediately without any training or help?;
- Haptics: what is the haptic feeling resulting from using the EZExFix?

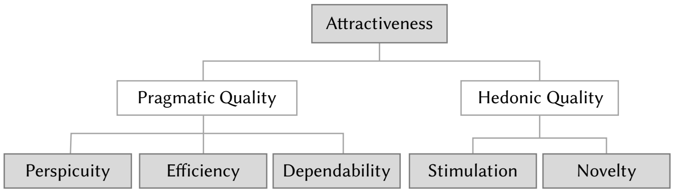


Fig. 2. Initial Scale structure of UEQ [22].

Attractiveness is an overall positive or negative impression of the product, while Perspicuity and Efficiency are hard aspects of the user experience representing the pragmatic quality of the device. Users typically perceive products with greater pragmatic quality as intuitive to use, efficient, and trustworthy.

UEQ+ was used to compare various designs of Playbook, a self-scheduling software used by astronauts [26]. Each scale is, in turn, decomposed into four subscales or items to be evaluated (*e.g.*, Attractiveness is decomposed into annoying *vs.* enjoyable, bad *vs.* good, unpleasant *vs.* pleasant, and unfriendly *vs.* friendly), each subscale being a differential scale with 7 points between items of each pair (*e.g.*, annoying ⟶ enjoyable). Each item is measured employing a 7-point Likert-type scale with answer categories "Strongly disagree" (=1) to "Strongly agree" (=7). UEQ+ is selected as an evaluation method because it is a modular and modern evaluation method where scales can be decided based on the artifact to evaluate and cover its user experience, not just its usability. UEQ+ is also straightforward and cost-effective to administer to participants. The number of participants required is still an open question [27]: recruiting (analog) astronauts is a challenging task as very few candidates are available. Furthermore, for some scales, a comparison of their values leads to an interpretation of five effect sizes [28]: bad, below average, above average, good, and excellent. Our within-subject study design has two dependent variables:

1. The Scale mean score, a real variable that measures the average score obtained on all items on each scale for each of the six sessions $S_i \in \{S_1, S_2, S_3, S_4, S_{EVA}, S_{Stress}\}$.
2. The Scale importance, a real variable that measures the average weight of importance of each scale for each of the six sessions $S_i \in \{S_1, S_2, S_3, S_4, S_{EVA}, S_{Stress}\}$.

Participants' answers are computed with the UEQ data analysis tool and interpreted as follows [22]: "it is extremely unlikely to observe values above +2 or below -2,..., the standard interpretation of the scale means is that values between -0.8 and 0.8 represent a neutral evaluation of the corresponding scale, values superior to 0.8 represent a positive evaluation, and values inferior to -0.8 represent a negative evaluation". Based on this interpretation, the results obtained for the multi-session evaluation are first discussed regarding the global results for all scales, then regarding each individual scale.

C. Inter-Scale Global Results and Discussion

1. Interrater Consistency

Table I reports Cronbach's coefficient α [29] computed to quantify the internal consistency, which expresses the extent to which the scale measurements remain consistent within a session or over subsequent sessions under identical or different conditions. A high value indicates that the answers of participants across items are consistent. When participants give a high value for one of the scale items, they are also likely to provide high values for the other items. The mean coefficient for all scales on all sessions is $\alpha = .66$, which suggests a global questionable consistency, but close to $\alpha = 0.7$, which is considered as an acceptable consistency. Trust ($\alpha = 0.91$), Intuitive use ($\alpha = 0.86$), and Usefulness ($\alpha = 0.80$) received the highest values, thereby indicating that these three scales were consistently assessed by participants.

Although other scales received reasonably good values, Haptics ($\alpha = 0.07$) received the lowest value, highlighting that participants did not assess this scale uniformly, probably because they belong to three different profiles. Depending on their knowledge, they assessed in different ways this scale, which seems to be more profile-dependent as opposed to the others. On the one hand, the diversity of these profile categories improves the representativeness of participants but, on the other hand, they tend to lower their consistency. Among all items,

TABLE I. Interrater Consistency. Cronbach's α: α ≥ 0.9 = Excellent (E), 0.9 > α ≥ 0.8 = Good (G), 0.8 > α ≥ 0.7 = Acceptable (A), 0.7> α ≥ 0.6 = Questionable (Q), 0.6>α ≥ 0.5= Poor (P), α < 0.5 = Unacceptable (U)

Scale	Cronbach's α (interpretation)						
	S_1	S_2	S_3	S_4	Mean	S_{EVA}	S_{Stress}
Attractiveness	0.46 (U)	0.79 (A)	0.87 (G)	0.10 (Q)	0.55 (P)	0.93 (E)	0.89 (G)
Efficiency	0.69 (Q)	0.71 (A)	0.38 (U)	0.88 (G)	0.66 (Q)	0.55 (P)	0.90 (E)
Perspicuity	0.37 (U)	0.78 (A)	0.88 (G)	0.69 (Q)	0.70 (A)	0.62 (Q)	0.90 (E)
Trust	0.95 (E)	0.95 (E)	0.93 (E)	0.81 (G)	0.91 (E)	0.73 (A)	0.91 (E)
Adaptability	0.64 (Q)	0.82 (G)	0.10 (U)	0.81 (G)	0.59 (P)	0.85 (G)	0.15 (U)
Usefulness	0.73 (A)	0.96 (E)	0.66 (Q)	0.83 (G)	0.80 (G)	-0.63 (U)	0.91 (E)
Intuitive Use	0.82 (G)	0.92 (E)	0.87 (G)	0.81 (G)	0.86 (G)	0.43 (U)	0.80 (G)
Haptics	0.21 (U)	0.42 (U)	0.26 (U)	-0.61 (U)	0.07 (U)	-0.01 (U)	-0.64 (U)
Mean	0.61 (Q)	0.79 (A)	0.69 (Q)	0.54 (P)	0.66 (Q)	0.43 (U)	0.60 (Q)

TABLE II. Interrater Reliability. Kendall's W: $W \leq 0.2$ = Poor (P), $0.21 \leq W \leq 0.4$ = Fair (F), $0.41 \leq W \leq 0.6$ = Moderate (M), $0.61 \leq W \leq 0.8$ = Good (G), and $0.81 \leq W \leq 1$=Very Good (V)

Scale	Kendall's W (p-value, interpretation)						
	S_1	S_2	S_3	S_4	Mean	S_{EVA}	S_{Stress}
Attractiveness	0.35 (0.096, F)	0.11 (0.54, P)	0.10 (0.59, P)	0.18 (0.35, P)	0.185 (P)	0.31 (0.12, F)	0.23 (0.24, F)
Efficiency	0.075 (0.71, P)	0.31 (0.12, F)	0.086 (0.67 P)	0.067 (0.75, P)	0.134 (P)	0.042 (0.82, P)	0.13 (0.49, P)
Perspicuity	0.51 (0.025, M)	0.33 (0.11, F)	0.36 (0.084, F)	0.31 (0.13, F)	0.38 (F)	0.44 (0.047, M)	0.33 (0.10, F)
Trust	0.15 (0.41, P)	0.15 (0.41, P)	0.15 (0.44, P)	0.15 (0.44, P)	0.15 (P)	0.19 (0.32, P)	0.061 (0.77, P)
Adaptability	0.16 (0.38, P)	0.21 (0.26, F)	0.046 (0.82, P)	0.053 (0.81, P)	0.18 (P)	0.22 (0.25, F)	0.078 (0.70, P)
Usefulness	0.30 (0.14, F)	0.15 (0.44, P)	0.27 (0.17, F)	0.11 (0.54, P)	0.21 (F)	0.21 (0.26, F)	0.13 (0.49, P)
Intuitive Use	0.42 (0.053, M)	0.067 (0.75, P)	0.50 (0.029, M)	0.25 (0.20, F)	0.31 (F)	0.37 (0.08, F)	0.19 (0.32, P)
Haptics	0.28 (0.16, F)	0.29 (0.15, F)	0.41 (0.059, M)	0.45 (0.043, M)	0.36 (F)	0.51 (0.026, M)	0.30 (0.14, F)
Mean	0.28 (F)	0.21 (F)	0.24 (F)	0.21 (F)	0.24 (F)	0.28 (F)	0.18 (P)

Item1 "Stable-Unstable" ($M = 1.83$), Item3 "Rough-Smooth" ($M = 0.83$), and Item2 "Unpleasant to touch-Pleasant to touch" ($M = 0.50$) were rather positively assessed while Item4 "Slippery-Smooth" ($M = 0.0$) was rated as null. Perhaps the label "Smooth" shared by two bipolar scales confused participants. Three inter-item correlations of this scale were negative, Corr(Item3, Item4) $= -0.27$, Corr(Item1, Item2) $= -0.10$, and more surprisingly Corr(Item1, Item4) $= -0.80$, thereby suggesting that participants did not understand the items in the same way as the low values for some items counterbalanced the high values of other items, which creates a null effect. We therefore re-computed Cronbach's coefficient with missing items to obtain: $\alpha_{Item1} = 0.38$, $\alpha_{Item2} = 0.71$, $\alpha_{Item3} = 0.92$, and $\alpha_{Item4} = 0.66$. Remo-ving Item4 has a positive impact in our case.

In general, S_1 started with a questionable mean value ($\alpha = 0.61$), then increased to acceptable in S_2 ($\alpha = 0.79$) to return to a questionable one in S_3 ($\alpha = 0.69$) to sum up finally with an almost acceptable mean value ($\alpha = 0.66$). Haptics also drags the average consistency down from an acceptable global value ($\alpha = 0.72$ without Haptics) to a questionable one ($\alpha = 0.66$ with Haptics). However, the S_{Stress} situation ($\alpha = 0.60$), although fairly close, is interpreted as questionable.

2. Interrater Reliability

Since performance assessment in essential to experiments, interrater consistency and reliability are two indices that are commonly used to ensure such scoring consistency [30]. Therefore, after computing and reported the interrater consistency, we compute and evaluate the interrater reliability. Table II reports Kendall's coefficient of concordance W [31], a measure of agreement among participants which is equal to 0 when there is no agreement among them and 1 when a total agreement exists. The scales receiving the highest values are Perspicuity ($W = 0.38$), Haptics ($W = 0.36$), Intuitive Use ($W = 0.31$), and Usefulness ($W = 0.21$), all interpreted as a fair agreement. All other scales received a limited agreement when $W \leq 0.2$. Similarly, all sessions benefit from a fair agreement ($W \geq 0.21$), including the mean overall coefficient. Since Kendall's coefficient is very strict and demanding in its value, a fair value was not considered a disadvantage in our case, given the heterogeneity of the participants' profiles.

3. Scale Mean Scores and Importance

Fig. 3 shows the mean scores for the eight scales evalu-ated, each time across the four sessions (see Appendix for the detailed histograms for the four sessions). As a reference, S_{EVA} and S_{Stress} are represented each time as a horizontal bar calibrated on the corresponding mean scores. The mean scores and scale importance are positive for all scales for all sessions in this figure, even the most critical ones, which is very encouraging, thus explaining why the vertical left axis is reduced to [−1 ...3] instead of [−3... +3]. Only six out of 32 mean scores are below the 0.8 threshold, thus locating them in the neutral zone while all others are located in the positive one. Although below this threshold, the mean scores are not very far away: *e.g.*, Efficiency received 0.75 and 0.79 for S_1 and S_2, respectively, Adaptability received 0.67 and 0.79 for S_1 and S_2, respectively, with the exception that S_2 received the lowest mean score 0.33 for Haptics of all scales on all sessions.

Furthermore, a Wilcoxon signed rank test was calculated for a single sample to test significant differences above the median for each subscale for each session to discover that all the means of each subscale for each session were above their respective medians (*e.g.*, Attractiveness for S_1 gave a highly significant difference, *score* = 3.10, p = 0.0007*** with a large effect size r = 0.63) with only one exception (*i.e.*, Attractiveness for S_2 gave *z-score* = 1.35, p = 0.09, n.s.).

Shapiro-Wilk and d'Agostino-Pearson tests of normality were computed to determine whether the scale and sub-scale data are normally distributed. These results advised the rejection of the null hypothesis for all data and concluded that the data were not normally

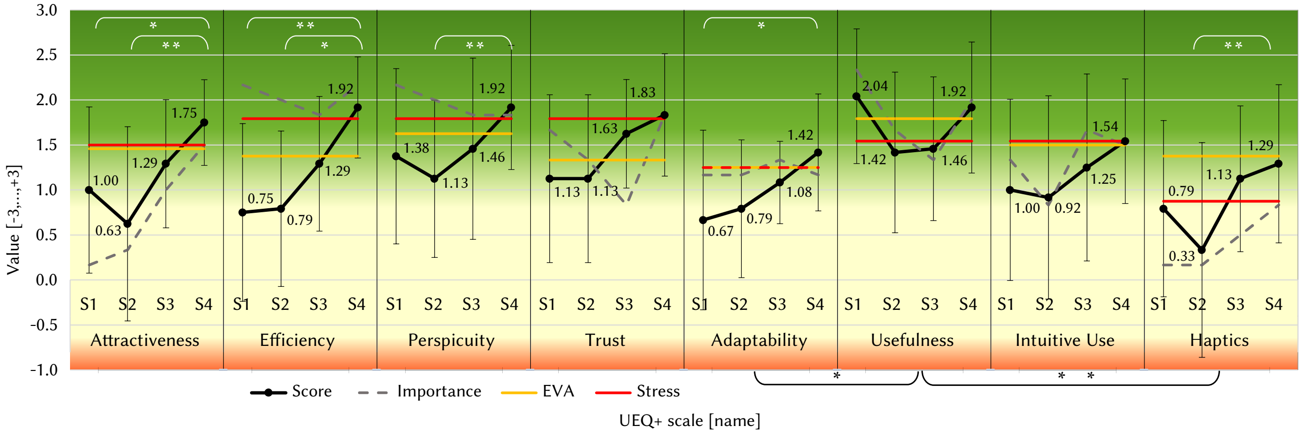


Fig. 3. Panel chart of the eight scales evaluated: mean scores (black bold straight lines) and scale importance (grey dotted lines) for sessions S_1 to S_4. Yellow lines show mean scores for SEVA and red lines show mean scores for SStress. Error bars show a confidence interval of 95% over mean scores. *: p < 0.05, **: p < 0.01.

distributed, since at least one test failed. Consequently, in the remainder of this paper, we will compute a non-parametric Friedman test of differences among repeated measures for all scales for all sessions with Dunn's test for multiple comparisons.

The first series of tests were carried out for the eight scales in all sessions and produced a Friedman statistic value of 22.61, which was significant (p = 0.002**): Usefulness is scored higher than Adaptability (R = −56, p = 0.0271*) and than Haptics (R = 66, p = 0.028**, see the bottom part of Fig. 3).

Overall, the mean scale scores start at S_1 with a rather positive value, then decrease or remain at the same level at S_2 to progressively increase again at S_3 and even more at S_4. For example, Attractiveness progresses as follows: it starts at $M_{S_1} = 1.00$, then decreases to $M_{S_2} = 0.63$, then increases to $M_{S_3} = 1.29$ to end at $M_{S_4} = 1.75$. All scale curves are globally increasing curves, *i.e.*, $\forall\, S_i \in \{S_1, S_2, S_3, S_4\}$: $M_{S_1} < M_{S_4}$ except for Usefulness but with very close values, *i.e.*, $M_{S_1} = 2.04 \geq M_{S_4} = 1.92$.

Contrary to an S-shaped performance curve that progressively increases until reaching a plateau or to an adoption curve that could decrease after the plateau, the eight factors that were continuously evaluated through a multi-session tend to follow a hype cycle curve. This type of curve starts with figures expressing a high expectancy in the device, then progressively decreases as the device is more frequently used in difficult and various conditions, to end up with a final increase to converge to a plateau expressing the final assessment of the device. For example, the important Haptics scale starts with a moderate mean score, then decreases and increases to end up with a more positive score. The importance is rated similarly.

At S_1, scales are sorted in decreasing order of their mean scores as follows: Usefulness (M = 2.04), Perspicuity (M = 1.38), Trust (M = 1.13), Attractiveness (M = 1.00), Intuitive Use (M = 1.00), Haptics (M = 0.79), Efficiency (M = 0.75), and Adaptability (M = 0.67). At S_4, this order remains mostly the same: Usefulness (M = 1.92), Perspicuity (M = 1.92), Efficiency (M = 1.92), Trust (M = 1.83), Attractiveness (M = 1.75), Intuitive Use (M = 1.54), Adaptability (M = 1.42), and Haptics (M = 1.29). Only the last two scales swapped their order, with Haptics slightly decreased but the Efficiency climbed up to the 3rd position. This result suggests that while the value of scale mean scores increased over sessions, participants tend to rate them in the same order except for the Efficiency that scales up throughout the training.

The mean scores of the scale for the stressful condition S_{Stress} (red lines in Fig. 3) are most of the time above the corresponding scores for the extra-vehicular condition S_{EVA}, except for Usefulness and Haptics. They even coincide for Attractiveness, Adaptability, and Intuitive Use, thus suggesting the real conditions in space may affect the user experience with respect to in-lab conditions, even with some stress imposed. Mean importance (grey dotted lines in Fig. 3) seems to follow the same hype cycle as scale mean scores, except for Usefulness.

4. Benchmarking of Scales

Schrepp *et al.* [28] mention more precise intervals to interpret some scales based on benchmarking performed on a set of evaluations. Each scale is decomposed into five intervals based on this benchmarking: bad, below average, above average, good, and excellent. Fig. 4 shows the interval in which the three concerned scales, *i.e.*, Attractiveness, Perspicuity, and Efficiency, are falling over the four sessions, from S_1 to S_4.

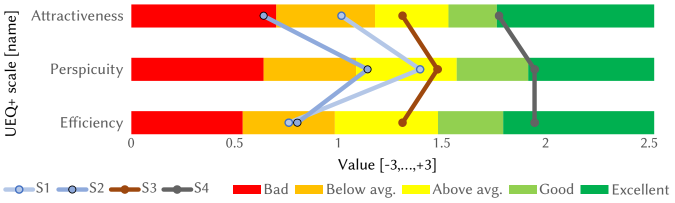


Fig. 4. Benchmarking of scales for all sessions S_i.

5. Key Performance Indicators

Another recent development of UEQ+ is the construction of the Key Performance Indicators (KPI) extension [32]. The KPI combines the subjectively perceived importance of user experience factors and the results of the UEQ+ into one figure. Fig. 5 shows the KPI for all sessions S_i. All mean KPIs are above the positive threshold of 1, except S_2 with a close value (M = 0.94, SD = 0.75). Indeed, S_1 (M = 1.11, SD = 0.37) initiates the multi-session that ends up with almost the same value in S_4 (M = 1.11, SD = 0.36). Interestingly, the KPI for the EVA (M = 1.47, SD = 0.23) and for the stress conditions (M = 1.54, SD = 0.54) are above the final value, thereby suggesting that participants were particularly attentive in expressing a higher performance in those critical conditions as opposed to normal ones. However, a non-parametric Kruskal-Wallis test revealed that there are no significant differences (H(5) = 5.79, α = 0.05, p = 0.33, *n.s.*) between the six KPIs.

D. Intra-Scale Results and Discussion

This section gathers all results for one scale at a time and consolidates them into a dedicated discussion considering scale mean scores (Fig. 3), and their mean importance (Fig. 6). Furthermore, this section concludes the individual discussion of each scale with the results of

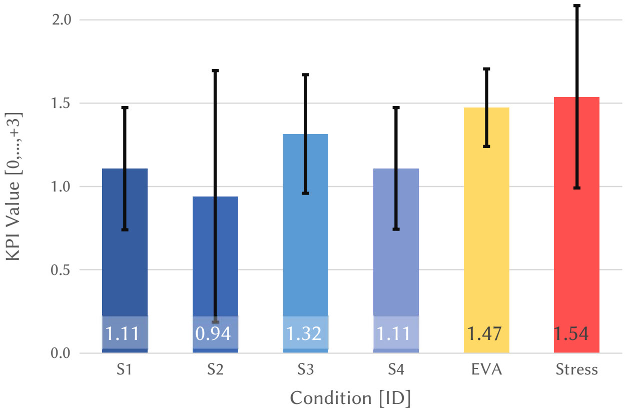


Fig. 5. Key Performance Indicators (KPI) for all sessions S_i. Error bars show the standard deviation.

TABLE III. Assignment Scales to IPAQuadrants According to the Two Methods. Each Quadrant Provides a Recommendation for Action: Q1="Keep Up the Good Work", Q2="Possible Overkill", Q3="Low Priority", Q4="Concentrate Here"

Scale	Scale Center		
	(0, 0)	Mean *S1*	Mean *S4*
Attractiveness	Q1	Q3	Q2
Efficiency	Q1	Q4	Q1
Perspicuity	Q1	Q1	Q1
Trust	Q1	Q1	Q1
Adaptability	Q1	Q3	Q3
Usefulness	Q1	Q1	Q1
Intuitive Use	Q1	Q3	Q3
Haptics	Q1	Q3	Q3

their Importance-Performance Analysis (IPA) [33]. This analysis aims to assign every scale to four different quadrants determined by two methods: (1) a differentiation by the coordinate origin at (0, 0), which is represented by a solid green line in Fig. 7; (2) a differentiation by the coordinate origin in the mean value of all scale values, which is represented by green dotted lines. Since we were interested in assessing the evolution of user experience of the EZExFix device over time, Fig. 7 shows the transition from the initial session S_1 to the final S_4: the X axis shows the performance computed as the scale mean score for the related session (S_1 and S_4, respectively) while the Y axis shows the mean importance for each scale. The dotted lines represent the average of all scale mean scores on X and the average of all mean importance on Y . The blue arrows show the transition from S_1 to S_4 for each single scale. Thus, each quadrant provides a recommendation for action for the respective scales, depending on its positioning. The plot is therefore divided into four quadrants [33]: Q1 = "Keep Up the Good Work" (top right quadrant when both the performance and the importance are above the corresponding mean value), Q2 = "Possible Overkill" (bottom right quadrant when the performance is above the corresponding mean value and the importance is below), Q3 = "Low Priority" (bottom left when both the performance and the importance are below the corresponding mean value), and Q4 = "Concentrate Here" (top left when the performance is below the mean value but the importance is above the mean value). These results are summarized in Table III. Note that all scales are located in Q1 with respect to the center of the scale at (0, 0) as they were all positive in terms of the mean values of the scale (Fig. 3) and the mean importance (Fig. 6).

Attractiveness. This scale is probably the most important among all scales assessed since it is supposed to capture "a user's general impression", one of the three dimensions of user experience. The perceived attractiveness of an artifact is considered to be the result of an averaging process of the perceived quality of the software with respect to the relevant aspects in a given usage scenario [23]. The mean score (±*SD*) of Attractiveness was 1.00 (±1.15) at baseline S_1, 0.63 (±1.35) at S_2, 1.29 (±0.89) at S_3 and 1.75 (±0.60) at S_4. All figures are above their respective importance values. The S_4 mean value is similarly above the critical conditions M_{EVA} and M_{Stress}, thus exceeding the expectations (Fig. 3). The hype cycle is even more revealing for this scale: a Friedman test ($F = 19.10$, $n = 4$) shows a significant difference ($p = 0.0003$***) between sessions. Post hoc analysis with Dunn's multiple comparison tests was performed with a Bonferroni correction applied, resulting in a significance level set at $p < 0.05$. A significant increase was observed between S_1 and S_4 ($R = -24.00$, $p = 0.0437$ *) and between S_2 and S_4 ($R = -28.50$, $p = 0.086$**). Furthermore, the standard deviation is progressively reduced as sessions progress: from $SD_{S1} = 1.15$ to $SD_{S4} = 0.60$. The overall good assessment is reinforced by a final 'excellent' position in benchmarking (Fig. 4) and a "Q2=Possible overkill" position (Table III). This should be mitigated by varying interrater consistency (limited on average, but good in critical conditions) and reliability (again limited on average, but fair in critical conditions), probably due to the small number of people having heterogeneous profiles.

Efficiency. This scale is considered to be a pragmatic quality of user experience and is "goal-directed" (its assessment is based on tasks that can be performed with the device) [28]. The mean score (±*SD*) of Efficiency was 0.75 (1.23) at baseline S_1, 0.79 (1.08) at S_2, 1.29 (0.93) at S_3 and 1.92 (0.70) at S_4. All figures are below their respective importance values. Typically, expectations are met when the mean values are equal to or greater than their importance. However, in this case, all values are highly positive, the difference between both values at S_4 is small and the S_4 mean value is still above the critical

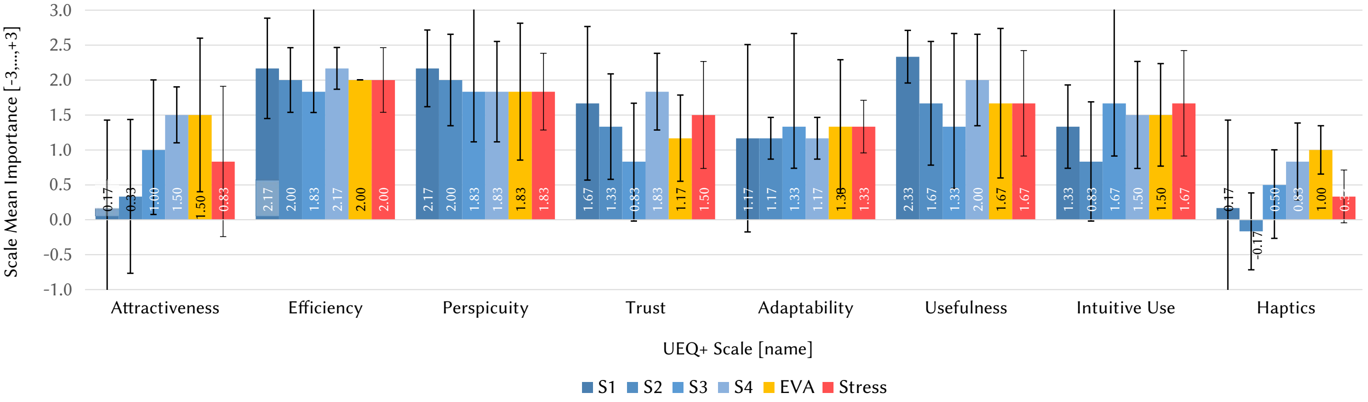


Fig. 6. Scale importance for sessions S_1 to S_4, S_{EVA}, and S_{Stress}. Error bars show a confidence interval of 95%.

conditions M_{EVA} and M_{Stress}, thus being anyway interpreted as a good assessment (Fig. 3). Similarly to Attractiveness, this scale also knows a significant increase between the initial session S_1 and the final session S_4 ($R = -32.00$, $p = 0.0021^{**}$, *z-score* $= 3.578$) and between S_2 and S_4 ($R = -26.50$, $p = 0.0183^{*}$, *z-score* $= 2.963$). Efficiency ends with the highest mean score of all scales ($M_{S_4} = 1.92$) and is interpreted as 'excellent' in the benchmarking (Fig. 4). Although this scale was located in the "Q4=Concentrate Here" quadrant during S_1, its evolution reaches the best quadrant "Q1=Keep up the Good Work" during the final session S_4 with the best position of all scales (Fig. 7). The interrater consistency is better than that of Attractiveness, but with similar reliability for the same reasons.

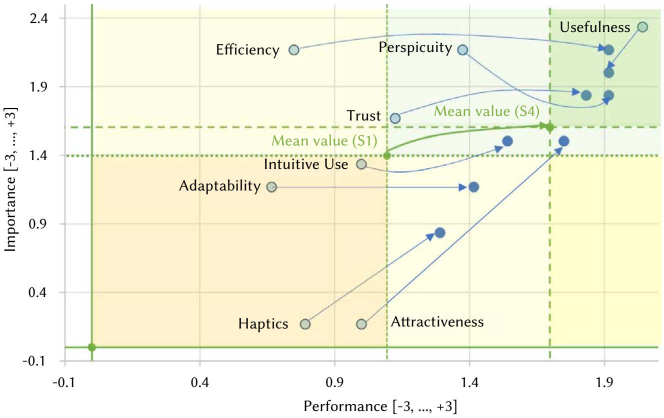


Fig. 7. Results of the IPA analysis: transition from S_1 to S_4.

Perspicuity. This scale expresses to what extent participants considered it easy to get familiar with the EZExFix device and to learn how to use it. Therefore, it is also considered a pragmatic quality that is an important part of the user experience and is "goal-directed" [28]. The mean score ($\pm SD$) of Perspicuity was 1.38 (1.22) at baseline S_1, 1.13 (1.09) at S_2, 1.46 (1.26) at S_3 and 1.92 (0.86) at S_4. The S_4 mean value is above its importance and critical conditions (both M_{EVA} and M_{Stress}) (Fig. 3). Several signs concur to conclude that this scale is very positively and rigorously assessed: its importance remains consistently estimated across sessions (Fig. 6), it knows a significant increase only between S_2 and S_4 ($R = -29.50$, $p = 0.0058^{**}$, *z-score* $= 3.298$) (Fig. 3), it is interpreted as 'excellent' in the benchmarking (Fig. 4), it consistently stays in the quadrant Q1="Keep Up the Good Work" during all sessions considered, and it is the third of all scales in this quadrant (Fig. 7).

Trust. This scale expresses the extent to which participants are confident in the use of the device, in its correct functioning, and, above all, that it will not harm them, which is crucial in the case of a limb fracture. The mean score ($\pm SD$) of Trust was 1.13 (1.17) at baseline S_1 and at S_2, 1.63 (0.75) at S_3 and 1.83 (0.85) at S_4. The S_4 mean value is nearly equal to its importance and M_{Stress}, and above M_{EVA} (Fig. 3). There were no significant differences among matched UEQ+ items of this scale between the different sessions (Fig. 3). It stays consistently in quadrant Q1 = "Keep Up the Good Work" during all sessions considered ending in the fourth place of all scales (Fig. 7), thereby suggesting that it was positively recognised, especially with excellent average consistency (the best among all scales).

Adaptability. This scale expresses the extent to which participants felt that the device could be adapted to a range of personal parameters, such as their own physical configuration, preference, or individual way of working. Although we did not take into account a factor of the participants' physical morphology, which could be considered in this scale, the participants did not express a negative opinion in this respect. The mean score ($\pm SD$) of Adaptability was 0.67 (1.25) at baseline S_1, 0.79 (0.96) at S_2, 1.08 (0.57) at S_3 and 1.42 (0.81) at S_4 (Fig. 3). This shows that Adaptability received the lowest, yet neutral, scores during S_1, but that slightly increases over sessions until a positive value ($M_{S_4} = 1.42$) is obtained. This scale knows a significant increase only between S_1 and S_4 ($R = -25.00$, $p = 0.03118^{*}$, *z-score* $= 2.795$) (Fig. 3). This scale received low mean importance, and participants agreed to rate this scale below the corresponding means of all scales. Adaptability remains in the Q3="Low Priority" during S_1 to S_4 (Fig. 7), thereby suggesting that any form of adaptation is not that important for the participants.

Usefulness. This scale expresses how useful the participants felt the device was in fixing a broken leg, which is already a critical situation, even though it was assessed under normal, stressful conditions without any participant actually having a limb with a fracture. The mean score ($\pm SD$) of Usefulness was 2.04 (0.93) at baseline S_1, 1.42 (1.11) at S_2, 1.46 (1.00) at S_3 and 1.92 (0.91) at S_4. All the mean values are close to their corresponding scale importance and end up above the critical conditions (both M_{EVA} and M_{Stress}) (Fig. 3). Among all scales, Usefulness received the highest mean scores both in S_1 and S_4, which makes this scale the most positively assessed. This is particularly important since the main goal of the device lies in its usefulness first, and in its user experience second. There were no significant differences among matched UEQ+ items of this scale between the different sessions. This scale remains uniformly located in Q1="Keep Up the Good Work" both during S_1 and S_4 sessions (Fig. 7).

Intuitive Use. This scale expresses the extent to which participants were able to manipulate the device in the task assigned to them with minimal use of any form of assistance or guidance. The mean score ($\pm SD$) of Intuitive Use was 1.00 (1.26) at baseline S_1, 0.92 (1.41) at S_2, 1.25 (1.30) at S_3 and 1.54 (0.87) at S_4 (Fig. 3). The values during S_1 and S_4 were below, respectively similar with their corresponding importance and under critical conditions (both M_{EVA} and M_{Stress}). Thus, the mean score of this scale is only aligned with that of M_{EVA} and M_{Stress} when the last session S_4 was reached. There were no significant differences between the UEQ+ items matched on this scale between the different sessions ($F = 5.432$, $p = 0.1427$, *n.s.*)(Fig. 3). This scale remains in Q3="Low Priority" during S1 and S4, thus suggesting that the participants really needed the familiarisation to properly operate the device and that some effort should be devoted to improving this aspect.

Haptics. This scale expresses the extent to which participants felt their touch when handling the device, which is probably the most important factor as the device is supposed to provide the user with a sense of physical touch that best fits the task, i.e.,setting a fracture. This is largely covered by the pins of the device, but also by their configuration and handling. The mean score ($\pm SD$) of Haptics was 0.79 (1.22) at baseline S_1, 0.33 (1.49) at S_2, 1.13 (1.01) at S_3 and 1.29 (1.10) at S_4 (Fig. 3). None of the participants had any previous experience with such a device, nor did they have any experience in treating a fracture in a mission as perilous as that which one might imagine in space, on the moon, or on another planet such as Mars. This scale knows a significant increase only between S_2 and S_4 ($R = -31.00$, $p = 0.0032^{**}$, *z-score* $= 3.466$) (Fig. 3). Of all the scales studied, this one had the lowest start with the highest rise to finish with a respectable value, but lower than the other scales. This scale continuously remains in the third quadrant Q3="Low Priority" during S_1 and S_4 (Fig. 7).

E. Scale Correlation Analysis

Laugwitz *et al.* [23] reported that the UEQ scales were statistically independent of each other, apart for Attractiveness, thus assuming that conclusions related to Q4="Concentrate here" and Q1="Keep up the Good Work" will generate the highest impact. To confirm or to disconfirm that scales are indeed independent of each other, we computed Pearson's ρ correlation coefficient between the eight scales

TABLE IV. Inter-Correlations of the UEQ+ Scales: Pearson's ρ Coefficient

	Attractiveness	Efficiency	Perspicuity	Trust	Adaptability	Usefulness	Intuitive Use	Haptics
Attractiveness	–	0.48	0.38	0.13	0.47	0.27	0.38	0.20
Efficiency	0.93	–	0.34	0.14	0.22	0.07	0.52	0.04
Perspicuity	0.98	0.93	–	0.14	0.43	0.11	0.30	0.13
Trust	0.93	0.97	0.84	–	0.16	0.03	0.32	0.14
Adaptability	0.89	0.99	0.87	0.97	–	0.16	0.11	0.31
Usefulness	0.43	0.18	0.52	0.07	0.05	–	0.25	0.31
Intuitive Use	0.98	0.99	0.96	0.97	0.96	0.28	–	0.04
Haptics	0.97	0.84	0.90	0.90	0.79	0.38	0.91	–

Note: The upper-right half shows the correlation based on raw data (N = 196), the lower-left half those of the means across the four sessions (N = 4). For overall correlation, Pearson coefficients greater than ρ = 0.10 are significant, for correlations across means, coefficients greater than ρ = 0.30 are significant.

selected in our study, once across the whole sample (6 participants × 8 scales × 4 items = 196) and once based on the mean scores of the four sessions S_1 to S_4 ($N = 4$). The results are shown in Table IV. Following the guidelines recommended by Cohen [34], who proposed to interpret correlations of $\rho = 0.10$ as small, $\rho = 0.30$ as medium, and $\rho = 0.50$ as large, and consistently with Schankin *et al.* [35], we only interpreted correlations of $\rho > 0.30$ as being practically significant. For correlations across scale mean scores, correlations of $\rho > 0.30$ were statistically significant. As noted by Laugwitz *et al.* [23], Attractiveness is correlated with all other scales (all $\rho > 0.43$ in the lower-left part of Table IV. Although the scales are supposed to be independent of each other [23], we observed significant correlations between some of them: between Efficiency and Trust ($\rho = 0.97$), Adaptability, Intuitive Use ($\rho = 0.99$); between Perspicuity and Intuitive Use ($\rho = 0.96$); between Trust and Adaptability, Intuitive Use ($\rho =_{0.97}$). That is, scales measuring pragmatic aspects of EZExFix were correlated as well as those scales measuring non-pragmatic aspects.

IV. Conclusion

This paper presented and discussed the results of a multi-session evaluation of the EZExFix, a haptic device to be used by astronauts to fix a tibial shaft fracture in future lunar and Mars missions based on eight factors assessed by participants through corresponding items, scales, and importance ratings. The eight factors were continuously and uniformly evaluated through a multi-session, suggesting a hype cycle curve. The shape of this curve justifies the need for evaluating the scales over multiple sessions to reach a representative value and positioning in the quadrants. In the end, four of eight scales are located in the first ideal quadrant, while three are estimated to have low priority *i.e.*, Intuitive Use, Adaptability and Haptics. While Attractiveness is located in Q2 in the final session, it is so close to Q1 that we consider it encouraging.

These encouraging results should be moderated by the limitations of the study. Only 6 analogue astronauts were involved in the study because only one crew was evaluated. Having this kind of experiment is quite challenging because these facilities are not easily accessible to the general research community. They had three different backgrounds, which improves their diversity but also reduces their interrater consistency and reliability. Cautions should be taken in interpreting Usefulness, which might be somewhat biased in the sense that not all raters are equally knowledgeable in human physiology and fracture repair.

However, the high positivity of all scales allows us to be confident about the potential transposition to astronauts in real conditions. While orthopedic surgery is always based on objective learning curves, it is necessary to take into consideration the subjective learning curve, especially when it comes to putting a surgical device in the hands of astronauts without advanced medical training. The hype cycle curve attests the positive progress of the subjective perception.

Indeed, as the EZExFix is a device to treat injured astronauts in space conditions, to enhance survival and mission success, its Usefulness is considered of primary importance and meets the expectations. The Trust in the EZExFix remained constant during all sessions which is also very important in a surgical learning curve, applying the principle "*primum non nocere*" for both the patient and the operator. The best evolution along the different sessions was the Efficiency (Q4⟶Q1) which is task and/or goal-oriented and refers to the ability to do something with the minimum amount of time, effort, cost, or resources required to achieve the desired result. In other words, this is exactly what is sought to solve problems under extreme conditions, whether in space, in developing countries or in war medicine on Earth.

Also according to the astronauts themselves, a crew medical officer could be an essential member of a Mars mission [36] as the long-duration spaceflight and the harsh Martian environment pose various medical challenges that require expert knowledge, skills, and ability (KSAOs concept [37]) to manage. Astronauts would have more confidence in a physician with 4-6 years of clinical experience, which may increase the difficulty of selection [36]. However, our study showed that only four surgeries in a fortnight were enough to significantly upgrade the subjective learning curve of astronauts. By confronting it with the objective learning curve, the EZExFix could be a veritable tool to increase the autonomy and self-confidence of non-medical astronauts during long-duration exploration missions as well as the cost-effectiveness of meeting KSAOs requirements.

Finally, the multi-session assessment of the UEQ+ questionnaire could be of great interest for the evaluation of the subjective learning curve of surgical residents due to its ability to capture a broad range of emotional and cognitive responses during the learning process and provide valuable insights for improving surgical training programs on Earth.

Acknowledgment

The authors would like to deeply thank the Mars Society, Dr. Shannon Rupert and the Mission Support staff to welcome the crew in the MDRS (Utah desert) for the opportunity to execute this research simulation in a Mars analogue environment. We also want to thank Dr. Kouamé Jean-Eric Kouassi for his PhD thesis about the creation of the new EZExFix, Pr. Benoît Lengelé and Pr. Catherine Behets for their day-to-day support in the realization of our multiple projects, and Lies Fievé and Christine de Ville de Goyet for their help to create artificial broken legs.

This work was supported by a F.S.R. Fund («Fonds Spéciaux de Recherche», Belgium, Ref. ADi/16568.2021) and a «Student Angel Fund» (Ref. 304907648 val-31.05.), both granted by M.A.R.S. UCLouvain crew 2022, as well as a F.N.R.S. Aspirant Fund («Fonds National de la Recherche Scientifique», Belgium, Ref. ID 40004991) and a second F.S.R. Fund (Ref. ADi/IC/13936.2020), both granted by Dr. Julie Manon.

Detailed Histograms

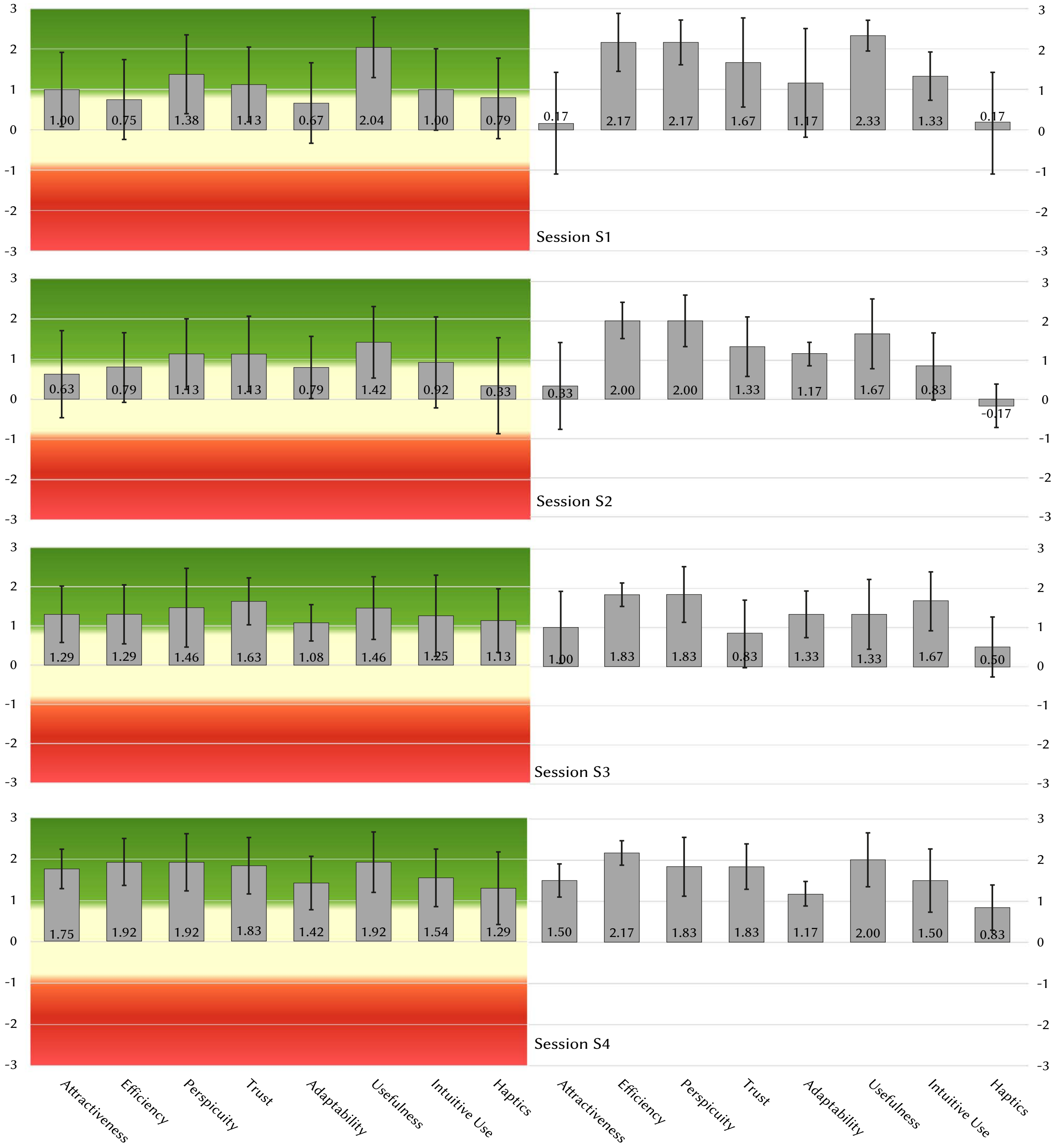


Fig. 8. Mean scores and scale importance of the four sessions. Error bars show a confidence interval of 95%.

References

[1] S. Brewster, “Haptic human-computer interaction,” in *Proceedings of the 4th Annual Conference of the ACM Special Interest Group on Computer-Human Interaction*, CHINZ ’03, New York, NY, USA, 2003, p. 3–4, Association for Computing Machinery.

[2] S. Villarreal-Narvaez, A. Sluÿters, J. Vanderdonckt, E. Mbaki Luzayisu, “Theoretically-defined vs. user- defined squeeze gestures,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, pp. 73–102, nov 2022, doi: 10.1145/3567805.

[3] V. Parthiban, P. Maes, Q. Sellier, A. Sluÿters, J. Vanderdonckt, “Gestural-vocal coordinated interaction on large displays,” in *Companion Proceedings of the ACM Symposium on Engineering Interactive Computing Systems*, EICS ’22 Companion, New York, NY, USA, 2022, p. 26–32, Association for Computing Machinery.

[4] Y. Fang, Y. Qiao, F. Zeng, K. Zhang, T. Zhao, “A human-in-the-loop haptic interaction with subjective evaluation,” *Frontiers in Virtual Reality*, vol. 3, 2022, doi: https://doi.org/10.3389/frvir.2022.949324.

[5] P. Xia, A. M. Lopes, M. T. Restivo, “Virtual reality and haptics for product assembly,” *International Journal of Online and Biomedical Engineering*, vol. 8, no. S1, pp. 12– 14, 2012, doi: 10.3991/ijoe.v8is1.1894.

[6] M. Haruna, M. Ogino, T. Koike-Akino, “Proposal and evaluation of visual haptics for manipulation of remote machine system,” *Frontiers in Robotics and AI, Section Smart Sensor Networks and Autonomy*, vol. 7, 2020, doi: 10.3389/frobt.2020.529040.

[7] J. Manon, C. Detrembleur, S. Van de Veyver, K. Tribak, O. Cornu, D. Putineanu, “Predictors of mechanical complications after intramedullary nailing of tibial fractures,” *Orthopaedics Traumatology: Surgery Research*, vol. 105, no. 3, pp. 523–527, 2019, doi: https://doi.org/10.1016/j.otsr.2019.01.015.

[8] J. Manon, C. Detrembleur, S. Van De Veyver, K. Tribak, O. Cornu, D. Putineanu, “Can infection be predicted after intramedullary nailing of tibial shaft fractures?,” *Acta Orthopædica Belgica*, vol. 86, pp. 313–319, 2020.

[9] J. Manon, C. Detrembleur, S. Van de Veyver, K. Tribak, O. Cornu, D. Putineanu, “Quels sont les facteurs prédictifs d'une complication mécanique après enclouage centromédullaire d'une fracture diaphysaire du tibia?,” *Revue de Chirurgie Orthopédique et Traumatologique*, vol. 105, no. 3, pp. 353–357, 2019, doi: 10.1016/j.rcot.2019.02.029.

[10] K. J.-E. Kouassi, J. Manon, L. Fonkoue, C. Detrembleur, O. Cornu, “Treatment of open tibia fractures in sub- saharan african countries: a systematic review,” *Acta Orthopaedica Belgica*, vol. 87, no. 1, pp. 85–92, 2021, doi: 10.52628/87.1.11.

[11] A. Terhorst, J. A. Dowling, “Terrestrial analogue research to support human performance on mars: A review and bibliographic analysis,” *Space: Science & Technology*, vol. 2022, 2022, doi: 10.34133/2022/9841785.

[12] J. Vanderdonckt, R. Vatavu, J. Manon, M. Saint- Guillain, P. Lefèvre, J. J. Márquez, “Might as well be on mars: Insights on the extraterrestrial applicability of interaction design frameworks from earth,” in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA 2024, Honolulu, HI, USA, May 11-16, 2024*, 2024, pp. 239:1–239:8, ACM.

[13] C. Salisbury, R. Gillespie, H. Z. Tan, F. Barbagli, J. Salisbury, “What you can't feel won't hurt you: Evaluating haptic hardware using a haptic contrast sensitivity function,” *IEEE Transactions on Haptics*, vol. 4, pp. 134–146, apr 2011, doi: 10.1109/TOH.2011.5.

[14] E. Samur, *Performance Metrics for Haptic Interfaces*. Springer Series on Touch and Haptic Systems, Springer, 2012.

[15] A. Hamam, A. E. Saddik, “Toward a mathematical model for quality of experience evaluation of haptic applications,” *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 12, pp. 3315–3322, 2013, doi: 10.1109/TIM.2013.2272859.

[16] R. Höver, M. D. Luca, M. Harders, “User-based evaluation of data-driven haptic rendering,” *ACM Transactions on Applied Perception*, vol. 8, nov 2010, doi: 10.1145/1857893.1857900.

[17] E. Samur, “Systematic evaluation methodology and performance metrics for haptic interfaces,” in *Proceedings of the IEEE World Haptics Conference*, WHC ’11, 2011, pp. 1–1.

[18] A. Ahmad, K. Andersson, U. Sellgren, M. Boegli, “Evaluation of friction models for haptic devices,” in *Proceedings of the Dynamic Systems and Control Conference*, vol. 2 of *Dynamic Systems and Control Conference*, 10 2013, p. V002T26A005.

[19] M. Saint-Guillain, J. Vanderdonckt, N. Burny, V. Pletser, T. Vaquero, S. Chien, A. Karl, J. Marquez, C. Wain, A. Comein, I. S. Casla, J. Jacobs, J. Meert, C. Chamart, S. Drouet, J. Manon, “Enabling astronaut self- scheduling using a robust advanced modelling and scheduling system: An assessment during a mars analogue mission,” *Advances in Space Research*, vol. 72, no. 4, pp. 1378–1398, 2023, doi: https://doi.org/10.1016/j.asr.2023.03.045.

[20] J. Manon, V. Pletser, M. Saint-Guillain, J. Vanderdonckt, C. Wain, J. Jacobs, A. Comein, S. Drouet, J. Meert, I. J. Sanchez Casla, O. Cartiaux, O. Cornu, “An easy-to-use external fixator for all hostile environments, from space to war medicine: Is it meant for everyonersquo;s hands?,” *Journal of Clinical Medicine*, vol. 12, no. 14, 2023, doi: 10.3390/jcm12144764.

[21] J. Manon, M. Saint-Guillain, V. Pletser, D. M. Buckland, L. Vico, W. Dobney, S. Baatout, C. Wain, J. Jacobs, A. Comein, S. Drouet, J. Meert, I. S. Casla, C. Chamart, J. Vanderdonckt, O. Cartiaux, O. Cornu, “Adequacy of in-mission training to treat tibial shaft fractures in mars analog testing,” *Scientific Reports*, vol. 13, 2023, doi: https://doi.org/10.1038/s41598-023-43878-1.

[22] M. Schrepp, J. Thomaschewski, “Design and validation of a framework for the creation of user experience questionnaires,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 7, pp. 88–95, 2019, doi: 10.9781/IJIMAI.2019.06.006.

[23] B. Laugwitz, T. Held, M. Schrepp, “Construction and evaluation of a user experience questionnaire,” in *Proceedings of the 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, HCI and Usability for Education and Work, USAB 2008*, vol. 5298 of *Lecture Notes in Computer Science*, 2008, pp. 63–76, Springer.

[24] A. Hinderks, M. Schrepp, M. Rauschenberger, J. Thomaschewski, “Reconstruction and validation of the UX factor trust for the user experience questionnaire plus (UEQ+),” in *Proceedings of the 19th International Conference on Web Information Systems and Technologies*, WEBIST 2023, 2023, pp. 319–329, SCITEPRESS.

[25] B. Boos, H. Brau, “Erweiterung des UEQ um die dimensionen akustik und haptik,” in *Proceedings of Usability Professionals*, UP 2017, 2017, Gesellschaft für Informatik e.V. / German UPA e.V.

[26] S. Shelat, J. A. Karasinski, E. E. Flynn-Evans, J. J. Marquez, “Evaluation of user experience of self- scheduling software for astronauts: Defining a satisfaction baseline,” in *Engineering Psychology and Cognitive Ergonomics*, Cham, 2022, pp. 433–445, Springer International Publishing.

[27] E. Schön, J. Hellmers, J. Thomaschewski, “Usability evaluation methods for special interest internet information services,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 2, no. 6, pp. 26–32, 2014, doi: 10.9781/IJIMAI.2014.263.

[28] M. Schrepp, A. Hinderks, J. Thomaschewski, “Construction of a benchmark for the user experience questionnaire (UEQ),” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 4, pp. 40–44, 2017, doi: 10.9781/IJIMAI.2017.445.

[29] L. J. Cronbach, “Coefficient alpha and the internal structure of tests,” *Psychometrika*, vol. 16, p. 297–334, 1951, doi: https://doi.org/10.1007/BF02310555.

[30] S.-C. Liao, E. Hunt, W. Chen, “Comparison between inter-rater reliability and inter-rater agreement in performance assessment,” *Annals of the Academy of Medicine, Singapore*, vol. 39, pp. 613–8, 08 2010, doi: 10.47102/annals-acadmedsg.V39N8p613.

[31] P. Legendre, “Species associations: the Kendall coefficient of concordance revisited,” *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 10, no. 226, 2005, doi: https://doi.org/10.1198/108571105X46642.

[32] A. Hinderks, M. Schrepp, F. J. D. Mayo, M. J. Escalona, J. Thomaschewski, “Developing a UX KPI based on the user experience questionnaire,” *Computers Standards & Interfaces*, vol. 65, pp. 38–44, 2019, doi: 10.1016/j.csi.2019.01.007.

[33] A. Hinderks, A. -L. Meiners, F. Mayo, J. Thomaschewski, “Interpreting the results from the user experience questionnaire (ueq) using importance-performance analysis (ipa),” in *Proceedings of the 15th International Conference on Web Information Systems and Technologies*, WEBIST 2019, Setubal, PRT, 2019, p. 388–395, SCITEPRESS, Science and Technology Publications, Lda.

[34] J. Cohen, *Statistical power analysis for the behavioral sciences*. New York, NY, USA: Routledge, 7 1988.

[35] A. Schankin, M. Budde, T. Riedel, M. Beigl, “Psychometric properties of the user experience questionnaire (ueq),” in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI ’22, New York, NY, USA, 2022, Association for Computing Machinery.

[36] I. S. Saluja, D. R. Williams, D. Woodard, J. Kaczorowski, B. Douglas, P. J. Scarpa, J.- M. Comtois, “Survey of astronaut opinions on medical

crewmembers for a mission to mars," *Acta Astronautica*, vol. 63, no. 5, pp. 586–593, 2008, doi: https://doi.org/10.1016/j.actaastro.2008.05.002.

[37] L. B. Landon, C. Rokholt, K. J. Slack, Y. Pecena, "Selecting astronauts for long-duration exploration missions: Considerations for team performance and functioning," *REACH*, vol. 5, pp. 33–56, 2017, doi: https://doi.org/10.1016/j.reach.2017.03.002.

Julie Manon

She graduated physiotherapist and Medical Doctor from the Université catholique de Louvain (UCLouvain, Belgium) in 2013 and 2018, respectively. She holds a master's degree in Orthopedic and Traumatology surgery and completed various university certificates (taping, basic and advanced trauma life support, master of animal experiments, physical and biological radiation protection, microsurgery, and statistics/data science). She obtained a Ph.D. dedicated to bone reconstruction (2024). Her research interests include understanding of massive bone grafts healing in a critical-size bone defect from fundamental mechanisms to preclinical studies. She is also interested in the fracture risk of astronauts and fixation possibilities to promote healing in a hostile spatial environment. For this purpose, she enrolled as an analog astronaut (health and safety officer) in a Mars simulation mission conducted at the Mars Desert Research Station (UT, USA) in 2022.

Jean Vanderdonckt

He received a master's in mathematics, a master's degree in computer science, and a Ph.D. in Sciences from the University of Namur, Belgium, in 1987, 1989, and 1997, respectively. He is a Full Professor at Louvain School of Management, UCLouvain, Belgium. His research interests include information systems, human-computer interaction (HCI), engineering interactive computing systems (EICS), intelligent user interfaces (IUI). He is Associate Editor of ACM Trans. on Interactive Intelligent Systems (TiiS), co-editor-in-chief of the Springer Series of Human- Computer Interaction and the Springer Briefs in Human-Computer Interaction. He is ACM Distinguished Scientist and IFIP Fellow.

Michael Saint-Guillain

He received a master's degree in computer science from UClouvain in 2013 and a Ph.D. in engineering science (UCLouvain, Belgium) and computer science (INSA-Lyon, France) in 2019, while studying artificial intelligence and combinatorial optimization under uncertainty. At that time, his research interests included logistics, operations management, and decision under uncertainty, initially applied to space exploration. Since 2019, he is CEO of Rombio, a university spinoff project, helping biotechnology and pharmaceutical manufacturing companies optimize their production, decisions, and assets. Furthermore, side research interests and contributions now include planning and scheduling in space, human-computer interface, and optimization techniques applied to medical particle physics.

Vladimir Pletser

He (Ph.D., MSc, MEng) is the Director of Space Training Operations at Blue Abyss, specializing in astronaut training. Previously, he was a senior Physicist-Engineer at European Space Agency (ESA) (1985–2016), managing ISS microgravity payloads and parabolic flight programs, logging a Guinness world record of 7,350 parabolas. He has trained astronauts, participated in Mars simulation missions, and served as a Visiting Professor at 25 universities worldwide. With over 650 publications, including books and journal articles, he is a member of several prestigious astronautical and scientific organizations.

Cyril Wain

He holds a master's degree in electrical engineering from UCLouvain, with a specialization in cryptography and telecommunication systems, as well as a master's degree in management sciences from Solvay (Belgium). Initially serving as a crew astronomer, he later became the commander of the Tharsis mission. He is currently a Belgian national trainee at ESA.

Jean Jacobs

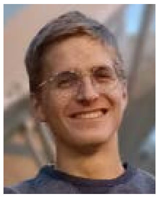
He obtained a master's degree in sciences, focusing on energy and environmental management (UCLouvain and Glasgow Caledonian University). He is a Ph.D. Candidate at the de Duve Institute (Belgium) and was the executive officer in the Tharsis mission.

Audrey Comein

She studied biological and biomedical sciences (Namur University, Belgium) and obtained her Ph.D. grade in 2025. She was enrolled twice for a manned mission (2020, 2021) and was the scientist in the Tharsis mission.

Sirga Drouet

She obtained a master's degree in biology (UCLouvain) and was the journalist in the Tharsis mission.



Julien Meert

He is a medical doctor at Cliniques Universitaires Saint-Luc, Université catholique de Louvain (UCLouvain, Belgium) where he is currently a "clinical assistant physician specialist candidate" (MACCS) in psychiatry and is interested in human psychological health and sleep. He played the role of the engineer in the Tharsis mission.

Ignacio Sanchez Casla

He holds a master's degree in mechanical engineering, from Ecole Polytechnique de Louvain (EPL), Université catholique de Louvain (UCLouvain, Belgium) and was enrolled as an astronomer in the Tharsis mission. He is currently a structural engineer at Societe Nationale de Construction Aerospatiale ("National Aerospace Construction Company").

Olivier Cartiaux

He holds a master's degree in electromechanical engineering (UCLouvain) with a specialization in mechatronics (2005). He further obtained a Ph.D. program focusing on computer and robotic assistance devices helping in orthopedic surgery (2010). After completing several Post-doctoral research, he is now the head of master's degree in health engineering (ECAM Brussels).

Olivier Cornu

He is head of the Orthopaedic and Trauma Surgery Department at the Cliniques Universitaires Saint-Luc UCL in Brussels and Professor of Anatomy and Physiology at UCLouvain. He has been deploying his expertise in the fields of musculoskeletal infections and tissue transplantation since 1996. His clinical practice is devoted to the management of infectious pathology of the musculoskeletal sector, to the reconstruction of large bone defects and revision joint replacement surgery. His research focuses on the study of the mechanical and biological properties of bone allografts, bone reconstruction, and implant-related infections, with an interest in treatments against bacterial biofilm. He also pursues numerous works oriented towards orthopedic and trauma care management in sub-Saharan Africa. He is an active member of several national and international scientific associations. He is also a member of the Royal Belgian Academy of Medicine and is Editor of the "Acta Orthopedica Belgica" Journal.

Evaluating Customer Segmentation Techniques in the Retail Sector

Nur Diyabi[1], Duygu Çakır[2], Ömer Melih Gül[1, 3]*, Tevfik Aytekin[1], Seifedine Kadry[4]

[1] Department of Computer Engineering, Bahcesehir University, Istanbul (Türkiye)
[2] Department of Software Engineering, Bahcesehir University, Istanbul (Türkiye)
[3] Informatics Institute, Istanbul Technical University, Istanbul (Türkiye)
[4] Department of Computer Science and Mathematics, Lebanese American University, Beirut (Lebanon)

* Corresponding author: omgul@itu.edu.tr

Received 26 January 2025 | Accepted 24 May 2025 | Early Access 28 May 2025

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

In the current competitive corporate landscape, understanding client preferences and adapting marketing strategies accordingly has become crucial. This study evaluates the effectiveness of four machine learning algorithms (K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Models (GMM), and Self-Organizing Maps (SOM)) for customer segmentation in the Turkish retail market. Two datasets were analyzed: a large-scale Turkish market sales dataset and a focused marketing campaign dataset. The research employed a comprehensive methodology encompassing data preparation, algorithm application, and performance evaluation using metrics such as the Calinski-Harabasz Index and Davies-Bouldin score. Results indicate that K-Means demonstrated superior performance in terms of interpretability and statistical validity. DBSCAN showed strengths in identifying non-spherical clusters, while GMM and SOM provided more granular segmentation. The findings offer actionable insights for Turkish retailers to optimize marketing strategies and enhance customer relationship management. This study contributes to the field of retail analytics by providing a methodological framework for evaluating customer segmentation techniques in specific market contexts.

Keywords

Clustering Algorithms, Customer Segmentation, Machine Learning, Retail Analysis, Unsupervised Learning.

DOI: 10.9781/ijimai.2025.05.001

I. Introduction

In the dynamic and competitive landscape of modern retail, understanding and effectively segmenting customers has transcended from being a mere advantage to becoming an absolute necessity. Customer segmentation, the meticulous process of grouping customers with similar characteristics and purchasing behaviors, has emerged as a critical strategy for businesses to navigate this complex terrain [1]. Traditional approaches to customer segmentation, while valuable, often fall short in capturing the patterns hidden within vast and complex datasets. The advent of machine learning techniques has opened new avenues for more sophisticated and accurate customer segmentation. These methods promise to uncover hidden patterns and insights that go beyond basic demographics, potentially revolutionizing how businesses understand and interact with their customers [2].

This paradigm shift is particularly evident in the Turkish retail market, where local supermarkets face the dual challenge of intense competition and rapidly evolving consumer preferences. The Turkish retail sector, characterized by its diversity and rapid growth, presents a unique context for studying customer segmentation. With major players like A101, BIM, CarrefourSA, and Migros dominating the field of supermarkets in Türkiye, the need for sophisticated customer insights has never been more pressing. These retailers are increasingly turning to data analytics to gain a competitive edge, with customer segmentation at the forefront of their strategies [3].

The primary goal of this study is to evaluate the effectiveness of four machine learning algorithms (K-Means clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Models (GMM), and Self-Organizing Maps (SOM)) for customer segmentation in the Turkish retail market. This evaluation involves comparing these algorithms across two datasets to identify their strengths and weaknesses in uncovering actionable customer segments. The study also assesses the algorithms using robust metrics and explores their practical implications for targeted marketing and customer relationship management, ultimately developing a framework for selecting the most suitable segmentation technique based on specific data and business needs.

Please cite this article as:
N. Diyabi, D. Çakır, Ö. M. Gül, T. Aytekin, S. Kadry. Evaluating Customer Segmentation Techniques in the Retail Sector, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 175-190, 2025, http://dx.doi.org/10.9781/ijimai.2025.05.001

This study contributes to the broader field of retail analytics by providing a methodological framework for evaluating customer segmentation techniques in specific market contexts. As businesses worldwide grapple with the challenges of data-driven decision-making, the findings offer insights that can inform strategy development and implementation across various retail environments. By identifying the most effective segmentation techniques for the Turkish retail market, local supermarkets can be equipped with the tools to make data-driven decisions about customer targeting strategies. This has far-reaching implications for enhancing customer satisfaction, optimizing marketing return of investment (ROI), tailoring product offerings, and ultimately fostering long-term customer loyalty in a highly competitive market.

In the following sections, the theoretical details of each segmentation technique will be analyzed, the methodology for comparison will be outlined, findings will be presented, and their implications for both practice and future research will be discussed. Through this comprehensive analysis, the understanding of customer segmentation in retail can be advanced, and actionable insights for businesses seeking to leverage data for competitive advantage in the dynamic world of modern retail can be provided.

This study is significant for advancing customer segmentation techniques by comparing advanced machine learning algorithms, with a focus on the Turkish retail market. It provides valuable insights for both academics and practitioners, offering practical implications for retailers to optimize their marketing strategies, enhance customer experiences, and improve decision-making through data-driven approaches. Additionally, the study contributes to the broader field of applied machine learning, with potential economic benefits for retail sector.

II. Literature Review

In customer segmentation, various methodologies have been developed across industries. This study focuses on four key approaches: Density-Based Spatial Clustering, Gaussian Mixture Models, Self-Organizing Maps, and K-means Clustering. These methods have been extensively studied and applied in customer segmentation, each offering distinct advantages and challenges.

The literature emphasizes the critical role of effective customer segmentation in enhancing marketing strategies, improving customer satisfaction, and driving business performance. It is stated that segmentation enables businesses to tailor offerings and communications to specific customer groups, leading to more efficient resource allocation and improved customer relationships [4].

Recent years have witnessed a shift towards machine learning-based approaches in customer segmentation. These methods have proven effective in identifying complex patterns within large datasets, a valuable capability in the current data-rich business environment. However, the effectiveness of these methods can vary depending on context and data characteristics. Mehrabi *et al.* [5] emphasize that factors such as data quality, algorithmic bias, and result interpretability must be carefully considered in real-world applications .

The following subsections will examine each approach in detail, focusing on their theoretical foundations, practical applications, and reported effectiveness in customer segmentation tasks. This review aims to provide a foundation for understanding the comparative analysis conducted in this study.

A. K-means Clustering Approach

K-means clustering has been widely recognized as a popular and effective unsupervised machine learning algorithm for customer segmentation. Its ability to cluster data points based on similarity without requiring labeled data has made it particularly useful in scenarios where customer labels are not readily available.

The application of K-means in customer segmentation has been extensively documented across various industries. In the retail sector, Kansal et al. [6] reported a case study where K-means was used to segment customers into four groups based on their shopping habits: high-value, medium-value, low-value, and at-risk customers. In the banking and financial services industry, Mohit [7] described the use of K-means to segment bank customers into three risk categories: low-risk, medium-risk, and high-risk, based on their risk profiles. Additionally, in telecommunications, Rungruang *et al.* [9] presented a study where K-means was employed to segment telecom customers into four groups based on usage patterns: heavy users, medium users, light users, and inactive users.

The process of applying K-means clustering for customer segmentation typically involves several key steps, as outlined in the literature. Determining the optimal number of clusters (k) is a crucial step that often involves trial and error, with various k values being assessed based on domain knowledge and business goals [7]. The initialization of centroids is also important, with advanced techniques like k-means++ being used to distribute centroids more evenly across the data [10]. The choice of distance metric, such as Euclidean or Manhattan distance, can significantly impact the clustering results [11]. Finally, the iterative process of updating centroids and reassigning data points continues until convergence is reached, indicating the successful identification of distinct customer segments [6].

While K-means has been widely applied and proven effective, several limitations have been identified in the literature. The algorithm's sensitivity to the initial placement of centroids can lead to suboptimal segmentation results [10]. Additionally, K-means assumes that clusters are spherical, which may not always align with real-world data distributions [11]. Furthermore, the requirement to pre-specify the number of clusters (k) beforehand can be challenging and may necessitate domain expertise or trial-and-error approaches [9].

Despite these limitations, K-means clustering is a popular choice for customer segmentation due to its simplicity, efficiency, and effectiveness in many practical cases.

B. Density-Based Spatial Clustering

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) has emerged as a powerful tool for customer segmentation, especially in scenarios where clusters have irregular shapes and varying densities. The algorithm's ability to detect clusters of any shape and its robustness against noise have been widely recognized in the literature [12]. DBSCAN operates by grouping together points that are closely packed in space, marking points that lie alone in low-density regions as outliers. This approach is particularly valuable in customer segmentation, where traditional centroid-based methods may fail to capture complex relationships between customers.

The process of DBSCAN clustering typically involves several key steps: constructing a neighborhood graph, where each node represents a data point and edges connect points within a specified distance (epsilon); identifying core points, which have at least a minimum number of points (MinPts) within their neighborhood; expanding clusters from core points to density-reachable points; and labeling points not belonging to any cluster as noise [13].

The effectiveness of DBSCAN in customer segmentation has been demonstrated across various industries. In a case study of a retail company, DBSCAN clustering resulted in the identification of three distinct customer groups [13]. These groups were characterized by different purchasing behaviors and demographic profiles, providing meaningful analysis for targeted marketing strategies.

One of the key advantages of DBSCAN, is its ability to handle outliers effectively [15]. In the context of customer segmentation, this translates to the ability to identify niche customer groups or unusual purchasing patterns that might be overlooked by other methods.

However, challenges associated with DBSCAN have also been identified in the literature. The selection of appropriate values for the epsilon and MinPts parameters can be critical to the algorithm's performance, as highlighted by Schubert *et al.* [16]. This selection often requires domain knowledge and can impact the resulting segmentation.

Despite these challenges, DBSCAN has been widely adopted for customer segmentation tasks, particularly in scenarios where the shape of clusters is not known *a priori*. Its ability to identify clusters of varying densities and shapes makes it a valuable tool in the increasingly complex landscape of customer behavior analysis.

C. Gaussian Mixture Model Customer Segmentation

Gaussian Mixture Models (GMMs) have been increasingly applied in customer segmentation due to their ability to model complex, multi-modal data distributions. As described by Scientific [17], GMMs model the data as a mixture of Gaussian distributions, with each distribution potentially representing a distinct customer group.

The application of GMMs in customer segmentation has been documented across various industries. In the retail sector, Zakrzewska and Murlewski [8] reported the use of GMMs to categorize retail customers into four segments: high-value, medium-value, low-value, and at-risk, based on purchasing habits, demographics, and other characteristics, where they utilized a hybrid GMM-fuzzy logic model to segment bank customers into three risk categories: low-risk, medium-risk, and high-risk, based on account activity and other variables.

The effectiveness of GMMs in customer segmentation has been attributed to their ability to capture complex, multi-dimensional relationships in customer data. Naga's study [17] on a dataset including customer age, demographics, gender, income, and purchase history reported an accuracy of 70% in customer segmentation using GMM.

However, several challenges associated with GMM-based segmentation have been noted in the literature. Determining the optimal number of clusters can be difficult, as selecting too many Gaussian components may lead to overfitting [17]. The computational cost of training GMMs on large datasets can be high, potentially limiting their use in real-time customer segmentation scenarios [18]. Additionally, the complexity of GMMs can make it challenging for businesses to interpret the resulting customer segments and translate them into actionable marketing strategies [19]. GMMs are also sensitive to the initial values of model parameters, which can affect performance [20]. Lastly, GMMs operate under the assumption that the data is generated from a mixture of normal distributions, an assumption that may not always hold in real-world scenarios [21].

Related to probabilistic clustering approaches, fuzzy clustering methods have also shown promise in customer segmentation applications. A recent study by Saadi et al. [22] demonstrated how fuzzy clustering can improve retrieval performance in case-based reasoning systems, suggesting potential synergies between fuzzy methods and customer segmentation tasks. While our study focuses on GMM, future research could explore the comparative performance of fuzzy clustering approaches in the Turkish retail context.

D. Self-Organizing Maps Clustering

Self-Organizing Maps (SOMs) [27], an unsupervised machine learning algorithm, have been widely applied in customer segmentation due to their ability to handle complex, high-dimensional data. The effectiveness of SOMs in visualizing and analyzing such data has been documented in studies [23].

The utility of SOMs in customer segmentation has been demonstrated in several case studies. Üstebey et al. presented a case study on airline passengers, where SOMs were used to segment customers based on attributes such as ticket type, fare type, travel date, and total fare paid [26]. The study resulted in the identification of four distinct customer groups, providing observations for targeted marketing strategies.

Key advantages of SOMs in customer segmentation, as highlighted in the literature, include their ability to visualize high-dimensional data by projecting it onto a lower-dimensional space while preserving relationships between data points, which aids in understanding complex customer behaviors. Additionally, SOMs facilitate feature extraction from complex datasets, potentially uncovering hidden patterns in customer behavior. Moreover, SOMs are capable of handling non-linear relationships within the data, making them particularly suitable for complex customer datasets.

However, several challenges associated with the use of SOMs have been identified. SOMs can be sensitive to the initial state of the algorithm, potentially leading to different segmentation results based on initialization [28]. Additionally, the computational cost of training SOMs on large datasets can be significant, which may limit their applicability in certain scenarios. Furthermore, the complexity of SOMs can sometimes make it difficult for businesses to interpret the resulting customer segments and translate them into actionable marketing strategies.

To address these challenges, various techniques have been proposed in the literature. For instance, Valova *et al.* [28] suggested using multiple initialization techniques and selecting the model that yields the best results. Liu *et al.* [23] and Lundberg & Lee [20] proposed methods to enhance the interpretability of SOMs, including feature selection and the application of visualization techniques.

III. Problem Definition

In the Turkish retail market, the need for sophisticated customer segmentation techniques has become increasingly apparent. The problem addressed in this study is the evaluation and comparison of various machine learning algorithms for customer segmentation, with a focus on their applicability and effectiveness in the Turkish retail sector.

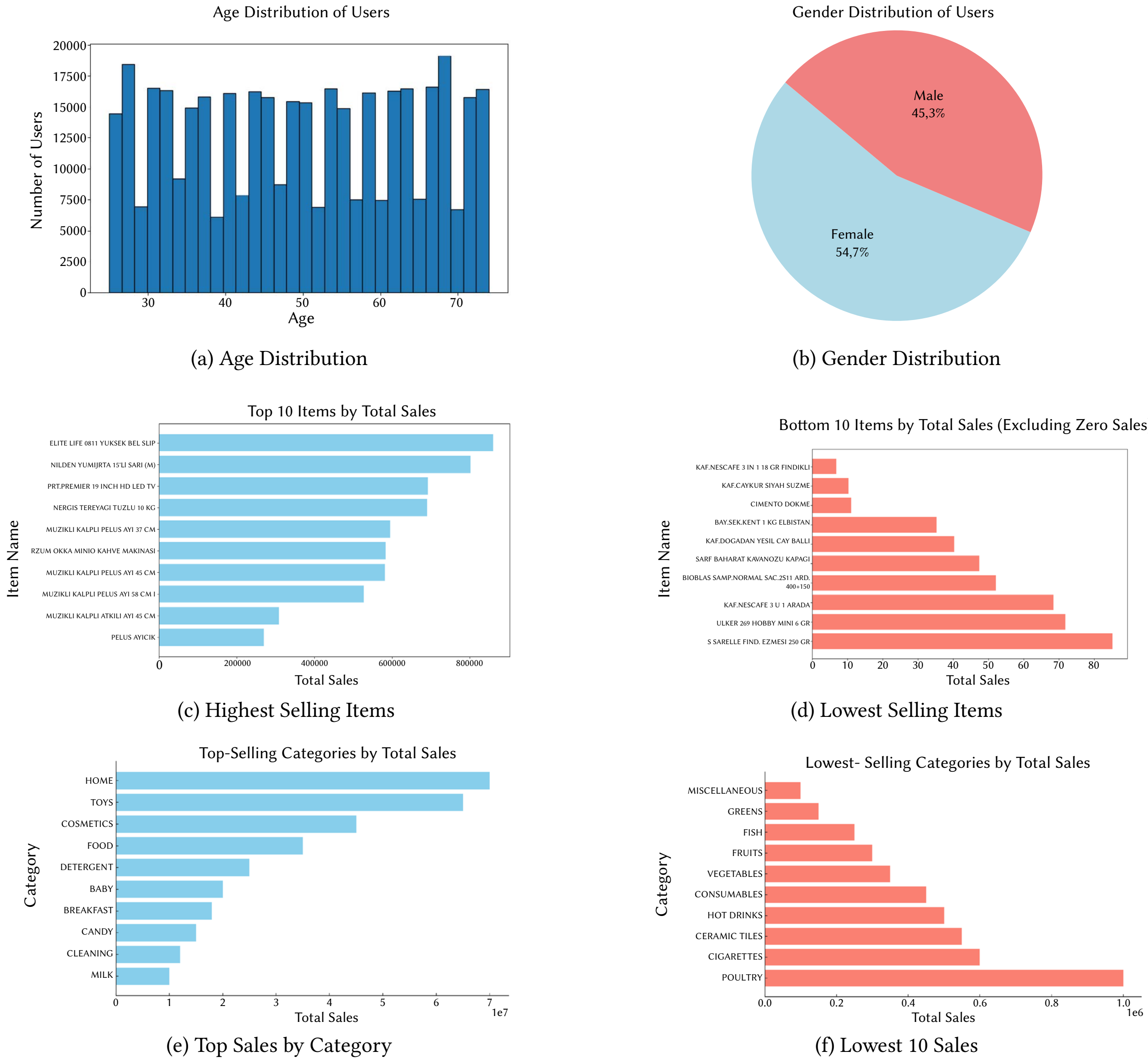
The primary challenge lies in determining which of the four selected algorithms is the most effective and actionable customer segmentation technique for the unique characteristics of the Turkish retail sector: K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Models (GMM), and Self-Organizing Maps (SOM). This problem is compounded by the diverse nature of available data and the specific characteristics of the Turkish market.

Two distinct datasets are utilized in this study to provide a comprehensive evaluation:

1. Turkish Market Sales Dataset: A large-scale Turkish market sales dataset comprising 10 million rows.
2. Marketing Campaign Dataset: A more focused marketing campaign dataset.

The use of these two datasets allows for the assessment of the algorithms' performance across different data scales and characteristics, which is crucial for understanding their practical applicability in various retail scenarios.

The problem addresses several critical aspects, including the identification of the most effective algorithms for segmenting customers based on their purchasing behavior and other relevant attributes. It also involves assessing the algorithms' ability to manage both large-scale



(a) Age Distribution

(b) Gender Distribution

(c) Highest Selling Items

(d) Lowest Selling Items

(e) Top Sales by Category

(f) Lowest 10 Sales

Fig. 1. Demographic, economic, and consumer behavior analysis on the TMS (first) dataset.

data and more focused datasets, while evaluating the interpretability and actionability of the resulting customer segments. Furthermore, the problem includes determining the computational efficiency and scalability of each algorithm, as well as assessing the robustness of the segmentation results across different data characteristics.

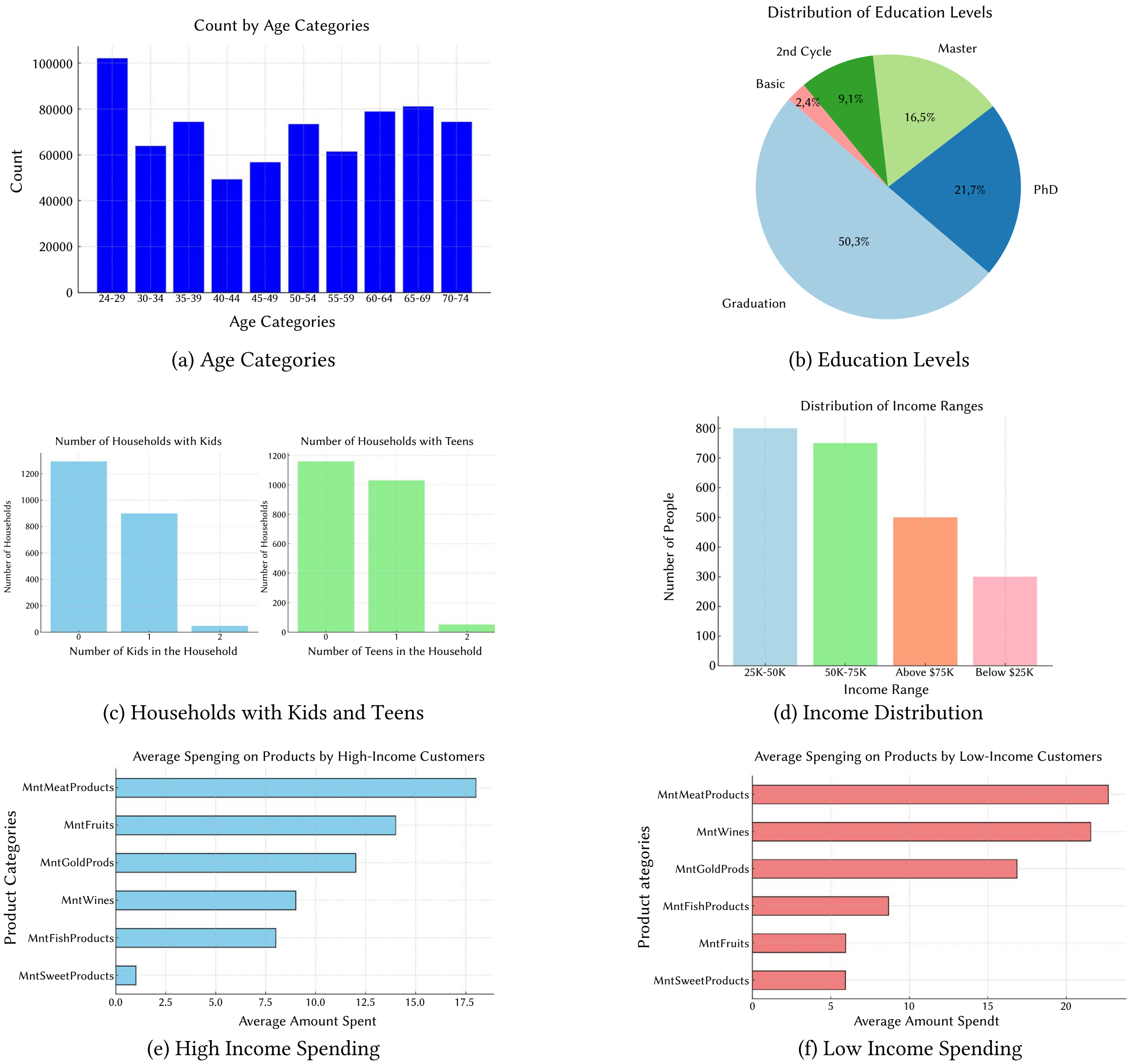
By addressing these aspects, this study aims to provide Turkish retailers to help find out the most suitable customer segmentation techniques for their specific needs and data characteristics. The ultimate goal is to enable more effective targeted marketing strategies, improved customer relationship management, and enhanced business decision-making in the Turkish retail sector.

A. Turkish Market Sales (TMS) Dataset

The first dataset utilized in this study is a comprehensive Turkish Market Sales Dataset, which provides a wealth of information about customer transactions at a local supermarket in Türkiye. This dataset is characterized by its large scale, comprising 10 million rows of transaction data [31].

This dataset is notable for its scale, consisting of 10 million rows of transaction data, which provides a substantial volume for evaluating algorithm performance on large-scale retail datasets. Each row represents an individual customer transaction, offering a detailed view of purchasing behavior at the transaction level. The dataset is also rich in features, encompassing various aspects of customer behavior and transaction characteristics. These features include customer demographics such as age and gender, product information like product category and brand, transaction details including purchase amount, date and time of purchase, payment method, and store location. The temporal aspect of the dataset, marked by the inclusion of transaction dates, enables the analysis of purchasing patterns over time, which is essential for understanding seasonal trends and the customer lifecycle. Additionally, the dataset is multi-dimensional, combining customer, product, and transaction data to provide a comprehensive view of customer behavior, facilitating complex segmentation analyses. Fig. 1 plots some of these insights from the dataset.

The use of this dataset presents several challenges and opportunities. The large scale of the dataset demands efficient data processing and analysis techniques, while the richness of available features requires careful consideration in selecting the most relevant ones for segmentation. The presence of categorical variables, such as product categories and payment methods, requires the application of appropriate pre-processing and encoding techniques.



(a) Age Categories

(b) Education Levels

(c) Households with Kids and Teens

(d) Income Distribution

(e) High Income Spending

(f) Low Income Spending

Fig. 2. Demographic, economic, and consumer behavior analysis on the MC (second) dataset.

The TMS dataset provides a realistic representation of the complexity and scale of data that large retailers in Türkiye might encounter, making it an excellent dataset for evaluating the scalability and effectiveness of different segmentation algorithms in the real-world retail scenario.

B. Marketing Campaign (MC) Dataset

The second dataset utilized in this study is the Marketing Campaign (MC) Dataset, which offers a focused examination of customer responses to various marketing initiatives [32]. Although this dataset is smaller in scale (2240 rows) compared to the Turkish Market Sales Dataset (10M rows), it provides a rich set of features that are particularly relevant for marketing campaign analysis and customer profiling. This dataset is more manageable in size, allowing for a detailed analysis of individual customer attributes and behaviors. Each entry represents an individual customer, giving a holistic view of customer characteristics, including demographics, behavioral data, purchase history, customer value metrics, and campaign response data. The dataset's diversity in features, such as birth year, education level, website visits, accepted deals, and specific purchase histories (e.g., wine and fruits), enriches the analysis. Additionally, it includes derived metrics like recency and customer tenure, which offer deeper insights into customer behavior and value.

The MC Dataset presents unique opportunities and challenges in the context of marketing analysis. Its multi-faceted customer profiles allow for the creation of detailed segments, facilitating more refined customer segmentation. The inclusion of campaign response data is particularly valuable for evaluating segmentation algorithms, enabling an assessment of how well these algorithms can identify customer groups with similar response patterns. The dataset's mix of numerical, categorical, and ordinal data requires careful pre-processing, making the handling of these varied data types a critical aspect of the analysis. Furthermore, the broad range of features necessitates a focused approach to determining feature importance, which is key to effective segmentation. The interaction between demographic factors and behavioral data within the dataset allows for a subtle exploration of how these elements define customer segments.

The MC Dataset complements the broader Turkish Market Sales Dataset by providing a detailed view of individual customers and their interactions with marketing campaigns. It enables the evaluation of segmentation algorithms in a context directly relevant to marketing strategy development and campaign optimization, thus playing a crucial role in this study's analysis. Demographic, economic, and consumer behavior analysis on the MC (second) dataset can be seen in Fig. 2.

While the two datasets provide complementary perspectives on Turkish retail customers, we acknowledge potential selection bias in our dataset selection. The Turkish Market Sales dataset may over-represent urban areas where such data collection is more feasible, while the Marketing Campaign dataset may have self-selection bias from customers who choose to participate in marketing programs. These limitations should be considered when generalizing our findings to the broader Turkish retail market.

IV. Methodology

In this study, a comprehensive methodological approach has been adopted to evaluate and compare four distinct machine learning algorithms for customer segmentation in the context of Turkish retail markets. The selected algorithms (K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Models (GMM), and Self-Organizing Maps (SOM)) were applied to two different datasets: a large-scale Turkish market sales dataset and a focused marketing campaign dataset. The selection of these four algorithms was based on their representation of different clustering paradigms commonly used in customer segmentation research. K-means represents centroid-based approaches (the most widely used baseline), DBSCAN represents density-based methods (suitable for non-spherical clusters), GMM represents probabilistic models (capable of handling overlapping clusters), and SOM represents neural network-based approaches (excellent for high-dimensional data visualization).

The methodology encompasses several key stages, including data pre-processing, feature selection, algorithm implementation, and evaluation. Each stage is carefully designed to ensure a rigorous and fair comparison of the algorithms' performance in customer segmentation tasks.

A. Proposed Approach

The proposed approach for this study involves a systematic comparison of four machine learning algorithms for customer segmentation. The primary method for initiating the segmentation process is K-means clustering, which serves as a benchmark against which the performance of other algorithms is measured. The approach can be outlined as follows:

1. ***Implementation of K-means Clustering:***

   The optimal number of clusters is determined using the Elbow method, followed by the application of K-means clustering on the pre-processed data. The resulting clusters are then analyzed and interpreted.

2. ***Application of Alternative Algorithms:***

   DBSCAN, GMM, and SOM are implemented on the same pre-processed data, with appropriate parameter tuning techniques employed for each algorithm.

3. ***Comparative Analysis and Context-Specific Evaluation:***

   The results from all algorithms are compared based on the performance metrics mentioned in subsection IV.G, and the interpretability and actionability of the resulting segments are assessed. Additionally, the computational efficiency and scalability of each algorithm are evaluated. The performance of each algorithm is assessed in the context of the Turkish retail market, and the applicability of the resulting segments to real-world marketing strategies is considered.

The proposed approach is designed to not only identify the most effective algorithm for customer segmentation but also to provide insights into the specific conditions under which each algorithm performs best. This information can be valuable for retailers in selecting the most appropriate segmentation technique based on their specific data characteristics and business objectives.

B. Data Preparation and Preprocessing

Data pre-processing is a crucial step in ensuring the quality and reliability of the customer segmentation results. For both the large-scale Turkish market sales dataset and the focused marketing campaign dataset, the following pre-processing steps were undertaken:

1. ***Data Cleaning***: Missing values were identified and handled appropriately, using techniques such as row deletions for the variables income and age.
2. ***Feature Engineering***: New features were created, such as deriving a "total spending" feature from individual transaction amounts in the sales dataset.
3. ***Encoding of Categorical Variables***: Categorical variables were encoded using appropriate techniques, with one-hot encoding applied to nominal categorical variables and ordinal encoding used for ordinal variables. For high-cardinality categorical variables, techniques such as frequency encoding or target encoding were considered to reduce dimensionality.
4. ***Feature Scaling***: Numerical features were scaled to ensure that all variables contributed equally to the analysis, with standardization (z-score normalization) applied to bring all numerical features to a common scale.
5. ***Data Type Conversion and Dimensionality Reduction***: Data types were converted as necessary to ensure compatibility with the chosen algorithms. For example, categorical variables were converted to numerical types for algorithms that require numerical inputs. Principal Component Analysis (PCA) has been applied to both datasets with n_components=3, to reduce the number of features while retaining most of the information.

C. Implementation of Clustering Algorithm

The core of the segmentation procedure is the k-means clustering algorithm. Based on their similarity, it repeatedly divides data points into a set number of clusters (k). Cluster centroids are initialized by the algorithm, either at random or with predetermined values. Our approach initialized clusters randomly and used a predetermined number of clusters obtained with the Elbow method. After that, each data point is assigned to the closest cluster centroid based on the Euclidean distance metric. Specifically, the algorithm assigns each data point to the cluster whose centroid has the minimum Euclidean distance from that point. The centroids are updated by taking the mean of all the data points in that cluster after the data points have been assigned. The centroids are updated and data points are assigned again until convergence is reached, at which point there is no more noticeable movement in the centroids.

K-means clustering served as the primary method for initiating the customer segmentation process. The approach involved (1) determining the optimal number of clusters using the Elbow method and Silhouette analysis,(2) initializing cluster centroids, and (3) assigning data points to the closest cluster centroid. Centroids were then (4) updated by taking the mean of all data points within each cluster, with steps three and four repeated until convergence was reached.

Determining the ideal cluster count (*k*) is essential for significant segmentation. The elbow method and silhouette analysis are two popular techniques.

Elbow method: Plotting the within-cluster sum of squares (WCSS) against the number of clusters (*k*) is the elbow method's method of analysis. The elbow point, where the WCSS begins to drop quickly and then stabilizes, is found to be the ideal number of clusters; this suggests that adding more clusters does not appreciably enhance the clustering result. We calculated WCSS for different k values and visualized it to identify the elbow point manually. Additionally, it leverages the function `KElbowVisualizer` in the library Yellowbrick, to automate the elbow method visualization, aiding in the selection of the most suitable number of clusters.

Silhouette analysis: determines how similar an object is to its own cluster versus other clusters. The silhouette score ranges between -1 and 1, with a higher score indicating that the object is well matched to its own cluster but poorly matched to neighboring clusters. The silhouette coefficient is calculated for each sample by taking the mean intra-cluster distance (a) and the mean nearest-cluster distance (b). A silhouette score close to 1 indicates that the sample is far away from the neighboring clusters; a score of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters; and a score of -1 indicates that the samples may have been assigned to the incorrect cluster. We calculated silhouette scores for various cluster counts and visualized them to manually determine the best number of clusters. Furthermore, we use the function `SilhouetteVisualizer` in the library Yellowbrick, to automate the silhouette method visualization, assisting in the selection of the optimal number of clusters. Silhouette Analysis indicates that the optimal value for *k* is 3.

D. Density Based Spatial Clustering of Applications With Noise (DBSCAN)

DBSCAN was implemented by considering two important features of every data point: a distance threshold (`epsilon`) and minimum neighbors (`MinPts`). The algorithm identifies clusters as high-density areas separated by low-density areas.

Consider a landscape with data points strewn all over it. High-density areas, such as busy city centers, are recognized by DBSCAN as clusters, which are made up of points that are close to one another and have lots of neighbors. Noisy areas are those with few data points, such as rural or suburban areas. DBSCAN looks at two important features of every data point. After fitting the model to the reduced-dimension customer data, cluster labels for each data point were extracted from the fitted model's labels attribute. These labels represent cluster membership or -1 for outliers. Then, the total number of clusters created, the number of outliers discovered, and the distribution of points within each cluster were examined.

E. Gaussian Mixture Model

Following the same steps as those employed in K-means clustering, including data preparation and cleaning, the GMM model was implemented. GMM provides granular segmentation by modeling the data as a mixture of Gaussian distributions, where each customer can have probabilistic membership across multiple clusters rather than hard assignments. This allows for more detailed understanding of customer segments, as some customers may exhibit characteristics of multiple segments. The code iterated through a range of cluster sizes, from 1 to 10, applying a GMM model to the data for each cluster size and computing the BIC score, a model selection metric. The number of clusters corresponding to the lowest BIC score was selected, indicating the most appropriate number of clusters for the data according to this metric. Subsequently, a GMM model was trained using the optimal number of clusters.

The algorithm estimates the parameters of these Gaussian distributions using the Expectation-Maximization (EM) algorithm [29], which iteratively refines the cluster assignments and distribution parameters until convergence. The trained model was then utilized to predict cluster labels for each customer data point, effectively assigning each customer to a specific segment based on their attributes. A new column labeled 'Cluster' was added to the data frame to include these cluster labels. The distribution of customers across segments was displayed by counting the occurrences of each cluster. To visualize the data, Principal Component Analysis (PCA) was employed, reducing the dimensionality of the data. Scatter plots were generated, with colors representing the various clusters, allowing for a visual inspection of how customers were classified based on their characteristics.

F. Self-Organizing Maps

The customer data is clustered using Self-Organizing Maps (SOM), an unsupervised learning technique for visualizing and analyzing high-dimensional data. Initially, missing values are handled, and numerical features are scaled to ensure consistency. The data is then converted into a format suitable for SOM analysis. The optimal SOM grid size is determined by comparing quantization errors across various grid sizes, with the appropriate grid size selected based on minimizing the quantization error as indicated by plotted results.

The SOM is trained on the data using the chosen grid size, and each data point is assigned to a cluster based on the winning neuron. The dataset is subsequently labeled into clusters for further analysis. The segmentation results are analyzed by computing the mean values of different attributes within each cluster, providing insights into distinct customer segments.

G. Evaluation Metrics

To assess the performance of the clustering algorithms, several evaluation metrics were employed:

- **Calinski-Harabasz Index**: Compares the ratio of between-cluster variance to the average within-cluster variance.
- **Davies-Bouldin Index:** Measures the ratio of within-cluster scatter to between-cluster separation.

While our study employs the state-of-the-art Calinski-Harabasz and Davies-Bouldin indices, recent advances in clustering validation have introduced new metrics such as the S-Divergence-Based Internal Clustering Validation Index [30], which provides an alternative approach to measuring cluster quality. Future work could benefit from incorporating these newer validation metrics.

V. Findings

The analysis of the Turkish market sales datasets using the four clustering algorithms resulted in several key findings, providing insights into customer segmentation within the Turkish retail market.

Two distinct datasets were employed in this study, each offering unique perspectives on customer behavior:

- A large-scale dataset comprising 1,000,000 rows and 28 columns, providing a comprehensive view of Turkish market sales. This dataset offered a broad spectrum of customer interactions and transactions, allowing for in-depth analysis of purchasing patterns across a wide customer base.
- A more focused marketing campaign dataset of 2240 rows and 29 columns, which, while smaller in scale, provided targeted information on customer responses to specific marketing initiatives. This dataset was particularly valuable for understanding the effectiveness of various marketing strategies and customer engagement levels.

A. Turkish Market Sales Dataset Results - Using K-Means Clustering

Initial analysis of the TMS dataset revealed key demographic distributions as shown in Fig. 1, and the correlation matrix of the dataset can be found in Fig. 3.

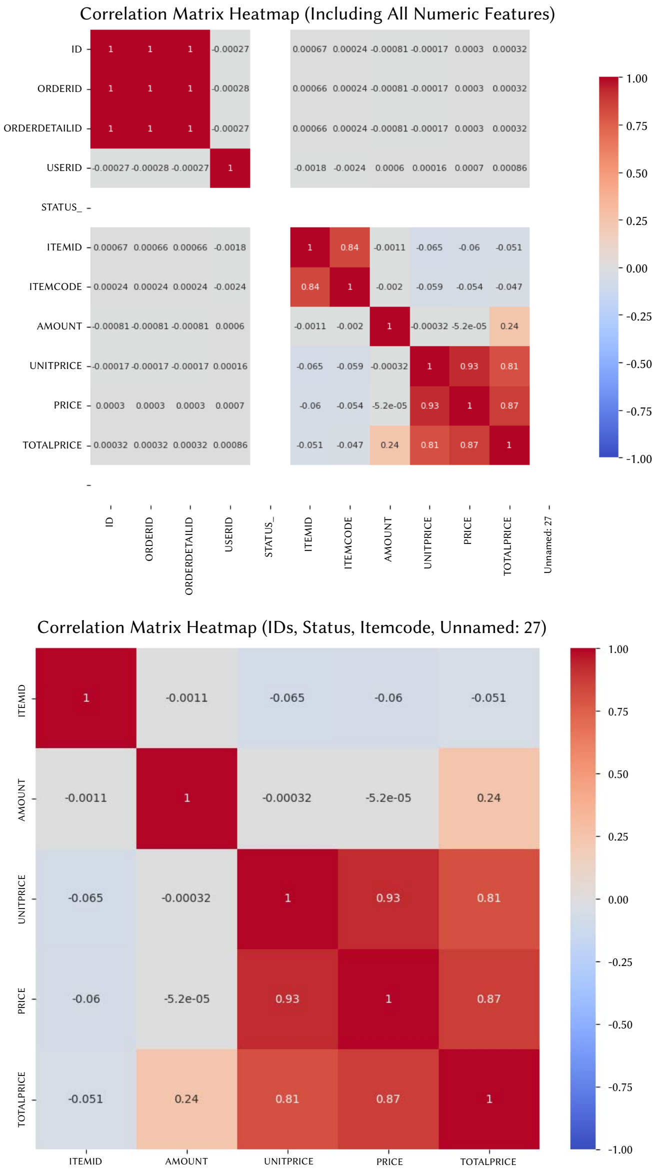


Fig. 3. Heatmap including (a) all numeric features, and (b) relevant features on the Turkish Market Sales Dataset.

The application of K-means clustering to the large-scale dataset revealed several important findings:

- Through the application of the Elbow method, it was determined that the optimal number of clusters for this dataset was 6 as illustrated in Fig. 4. This suggests that the Turkish retail market can be effectively segmented into six distinct customer groups, each with unique characteristics and behaviors (Fig. 5).
- To visualize these clusters, Principal Component Analysis (PCA) was employed. This technique allowed for the reduction of the high-dimensional data into a more manageable form, revealing clear and distinct customer segments. The visualization highlighted the separation between these segments, providing a clear picture of the market structure (Fig. 6).

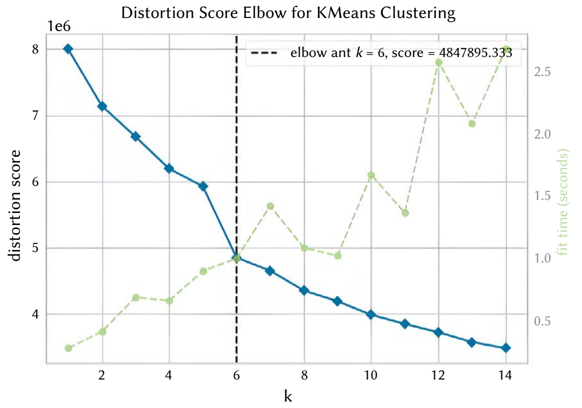


Fig. 4. Optimal cluster count detection using the Elbow method on the Turkish Market Sales Dataset.

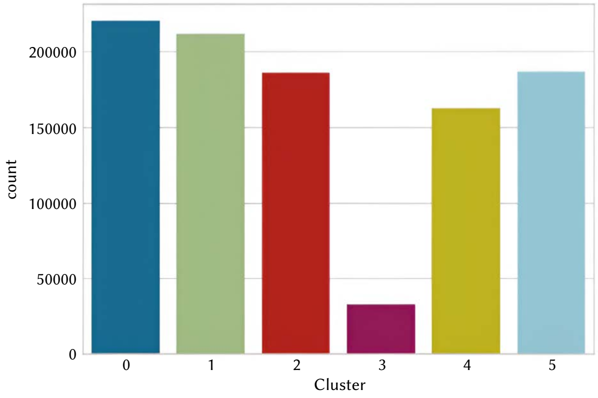


Fig. 5. Cluster analysis on the Turkish Market Sales dataset.

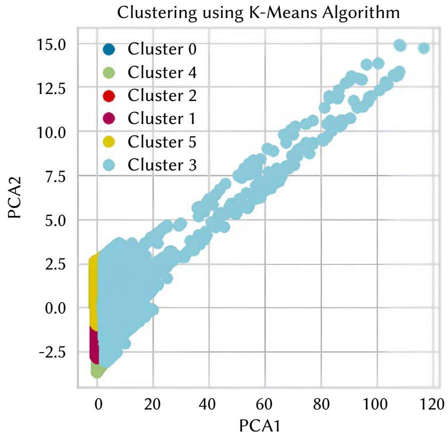


Fig. 6. PCA Dimension Extraction on the Turkish Market Sales dataset.

- The quality of the clustering was assessed using the Davies-Bouldin score, which was calculated to be less than 2 (1.6020). This low score is indicative of well-separated clusters, suggesting that the identified customer segments are distinctly different from one another. This clear separation is crucial for developing targeted marketing strategies for each segment.

Titles, labels, and a legend were added to the plot to ensure clarity and understanding. The visualization, created through the combined efforts of K-Means and PCA, offers valuable insights into the underlying structure of the data. The distribution of data points within each cluster can be observed, revealing potential groupings and unique characteristics. Subsequently, the original data was merged with the cluster labels assigned by K-means, resulting in a new DataFrame that incorporates these cluster labels. This enhancement allows for the analysis of features within each cluster, comparison of characteristics across groups, and a deeper understanding of the data's structure. It effectively tags each data point with a *group membership*, facilitating further exploration.

B. Marketing Campaign Dataset Results

Initial analysis of the Marketing Campaign (MC) dataset revealed key demographic distributions as shown in Fig. 2 and the correlation matrix of the dataset can be found in Fig. 7. The analysis of the focused marketing campaign dataset using various clustering algorithms provided detailed insights into customer segmentation.

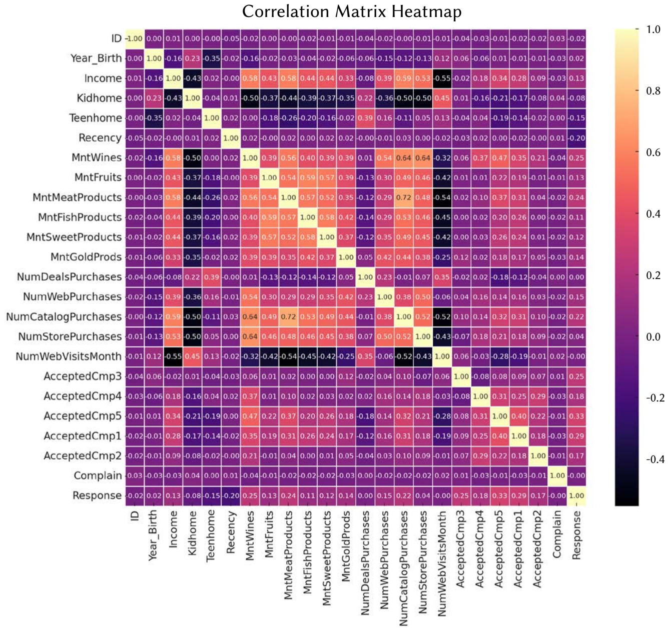


Fig. 7. Correlation matrix of the Marketing Campaign dataset.

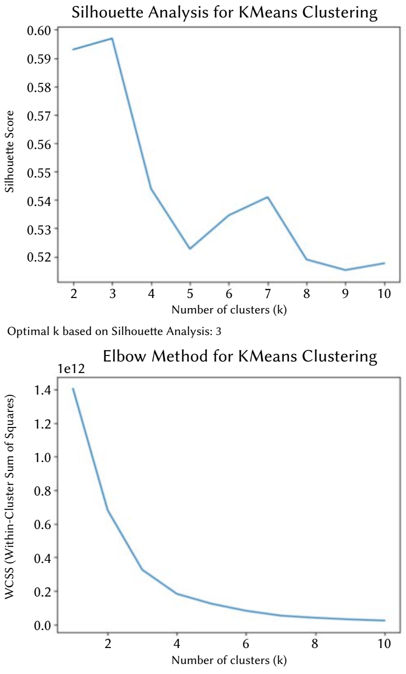


Fig. 8. Silhouette Analysis and Optimal Cluster Count Detection using the Elbow Method on the Marketing Campaign Dataset.

A Silhouette analysis and the Elbow method were performed, which identified the optimal number of clusters as three. This result assisted in determining the appropriate value for k, as illustrated in Fig. 8.

1. K-Means

The K-means algorithm identified three distinct clusters within the dataset, indicating three primary customer segments in the context of marketing campaign responses, as shown in Fig. 9. Notably, Cluster 1 emerged with the highest customer count, suggesting it as a dominant segment that could be a key target for marketing efforts, as illustrated in Fig. 10. Significant differences were observed across the clusters in terms of income levels, frequency of website purchases, and responsiveness to deal purchases, offering valuable insights for tailoring marketing strategies to the specific preferences and behaviors of each segment.

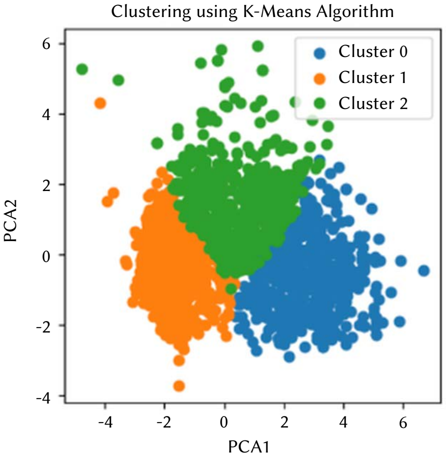


Fig. 9. PCA Dimension Extraction on the Marketing Campaign dataset.

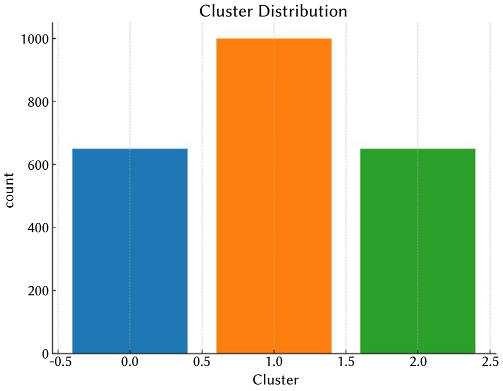
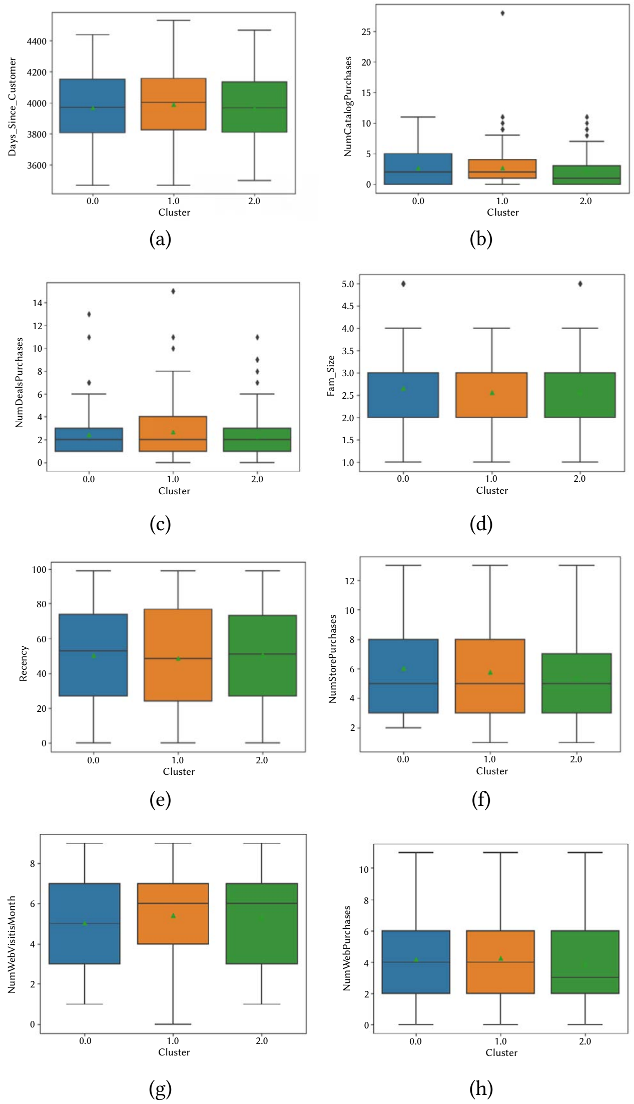


Fig. 10. Cluster analysis on the Marketing Campaign dataset.

Each cluster was visualised according to the results for each feature, as presented in Fig. 11, which provides valuable insights into the characteristics of the clusters. (a) presents the number of days since a member became a customer, identifying Clusters 0 and 2 as the oldest and most loyal customer groups. (b) highlights the number of purchases made through the catalogue, where Cluster 0 exhibited the strongest response, corresponding to the second cluster in the DBSCAN clustering approach Fig. 12. (c) shows the number of purchases made through supermarket deals, indicating that Clusters 1 and 2 are more responsive to deals, according to K-means results. (d) illustrates the family size of customers within each cluster, showing that most clusters consist of families with 2 or 3 members, indicating that family size was not a significant differentiating factor for clustering. (e) demonstrates the recency of purchases, with cluster 1 having the highest recency. (f) shows the number of in-store purchases, with Clusters 0 and 1 showing the highest purchase rates, emphasising the distinct purchasing behaviours of each cluster as identified by K-means. (g) displays the number of website visits, which corresponds to Cluster 9 in the GMM method shown in Fig. 13, providing insights into the patterns of website usage across clusters, which could inform future marketing strategies. (h) illustrates the number of purchases made through the website, revealing that Clusters 0 and 1 purchase from the website more frequently than Cluster 2, a pattern that is consistent with the K-means method but not clearly defined in GMM V.B.3 and SOM V.B.4 clustering algorithms.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 11. Box-plot analyses on the Marketing Campaign dataset. Details are in section 1, under K-means.

2. DBSCAN

The DBSCAN algorithm, known for its ability to identify clusters of arbitrary shape, revealed three main clusters and additionally identified outliers. This suggests the presence of niche customer groups that might be overlooked by other methods. Across these clusters, notable variations were observed in spending patterns, family size, and responsiveness to marketing campaigns. These insights offer a more detailed and subtle understanding of customer behavior, potentially uncovering unique market segments.

DBSCAN clustering results across the Marketing Campaign dataset can be found in Fig. 12. DBSCAN produced a relatively better Calinski-Harabasz score compared to other methods, suggesting a favorable balance between cluster density and separation.

The customer clusters were analyzed across key attributes, including spending patterns, family size, and purchasing behaviors. Total spending, shown in (a), reveals expenditure differences, with Cluster 0 as the highest spender. Family size (b), shows that most clusters have 2 to 3 members, with Cluster 0 having the largest families, indicating DBSCAN's consideration of this factor. Deal responsiveness in (c) highlights Cluster 2 as the most responsive, while Cluster 0 is the least, demonstrating the influence of deal purchases on clustering.

Store purchases, illustrated in (d), are highest in Cluster 2, consistent with GMM V.B.3 and SOM V.B.4 results. Website visits, shown in (e), indicate minimal differences, with Cluster 0 leading. Catalog purchases, in (f), also peak in Cluster 0, aligning with DBSCAN's focus on purchasing behavior.

Website purchases, shown in (g), place Cluster 2 as active online spenders. Age distribution in (h) shows Cluster 2 as middle-aged, while Cluster 1 is older. Cluster 2 also leads in accepted campaigns (i), reflecting their engagement. Income distribution, shown in (j), identifies Cluster 0 with the lowest income. Recency of purchases, in (k), shows Cluster 1 as the most recent buyers, and tenure (l), suggests Cluster 2 has the longest customer relationship. DBSCAN's clustering primarily focused on purchasing behavior while only slightly considering attributes like tenure.

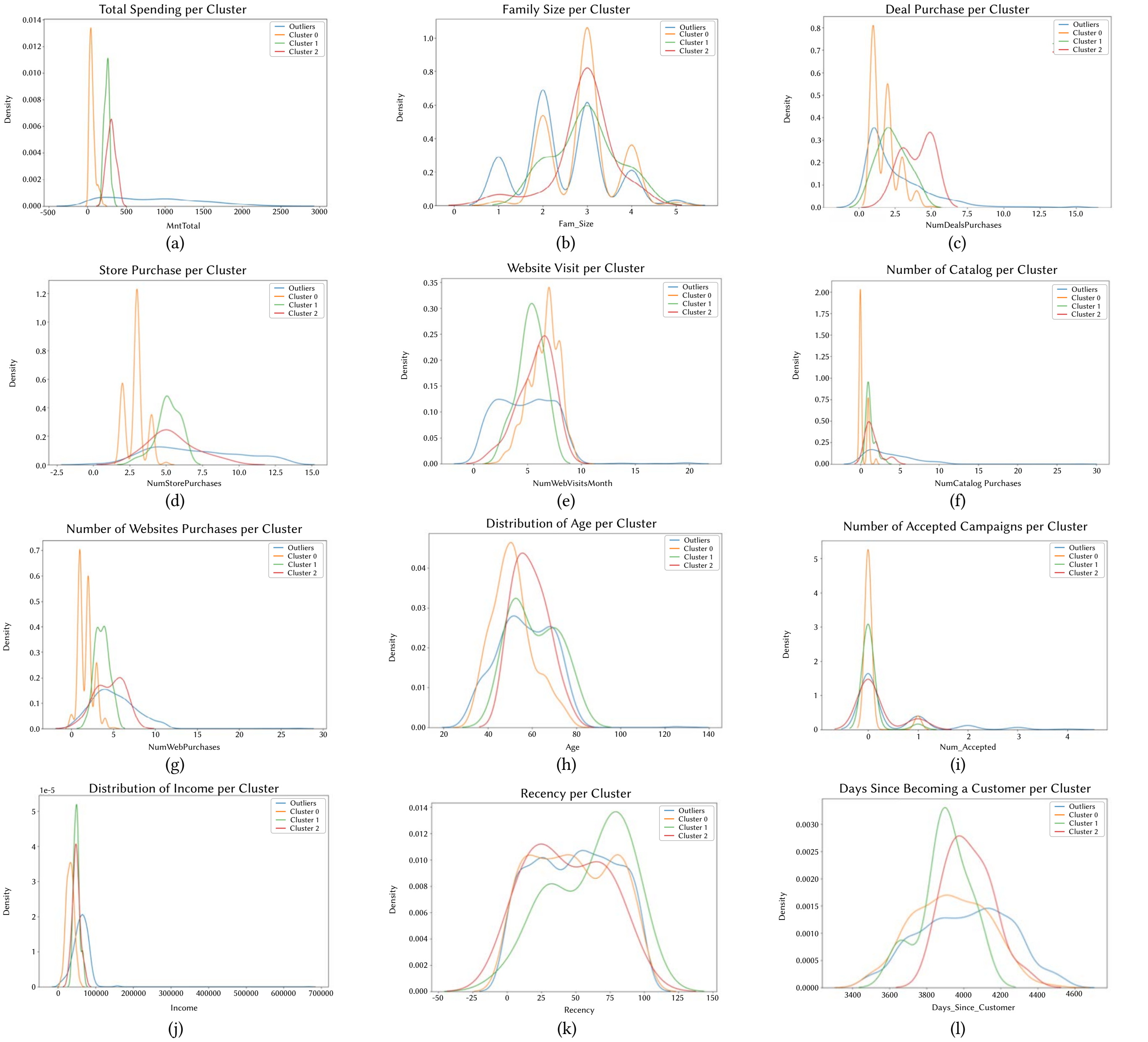
3. GMM

The Gaussian Mixture Model approach identified 10 distinct clusters, providing a more granular segmentation of the customer base. Each cluster exhibited varying characteristics in terms of spending habits, family size, and purchasing behavior, offering a highly detailed view of customer segments. Fig. 13 contains the results of each cluster analysis using the GMM clustering method according to the features.

The analysis and visualization of the ten customer clusters revealed distinct spending patterns, demographic characteristics, and purchasing behaviors. Cluster 6 showed the highest spending rate, identifying its members as the most active spenders, though spending varied within the cluster (a). Family sizes ranged mainly between 2 to 3 members, with Cluster 0 having larger families averaging 4 members, while Cluster 7 averaged 2 members (b). Cluster 0 also had the highest deal purchase response, marking these customers as the most receptive to promotions (c). Store purchases were highest in Cluster 1, while Cluster 7 had the least activity, including a group with zero store purchases (d). Cluster 0 led in website visits, aligning with its high deal purchase rate (e), and catalog purchases were strong in Clusters 3 and 6 (f). Cluster 1 dominated in website purchases, highlighting GMM's effectiveness in defining clusters based on purchase behaviors (g). Age distribution showed Cluster 2 as the oldest group, while Clusters 9 and 7 were the youngest (h). Campaign acceptance was highest in Cluster 6 (i), and Cluster 2 had the highest income, making it the wealthiest group (j). Cluster 4 had the most recent purchases (k), and Cluster 0 had the longest customer tenure (l), marking them as the oldest customer segment within the supermarket.

4. SOM

The Self-Organizing Maps technique resulted in the identification of 17 clusters, the highest number among all methods used. Each of these clusters presented unique characteristics based on factors such as age, income, and purchasing patterns. While this high number of clusters provides extremely detailed segmentation, it may present challenges in terms of practical application in marketing strategies. Cluster 1: Customers in this cluster tend to have an average family



(a) (b) (c)

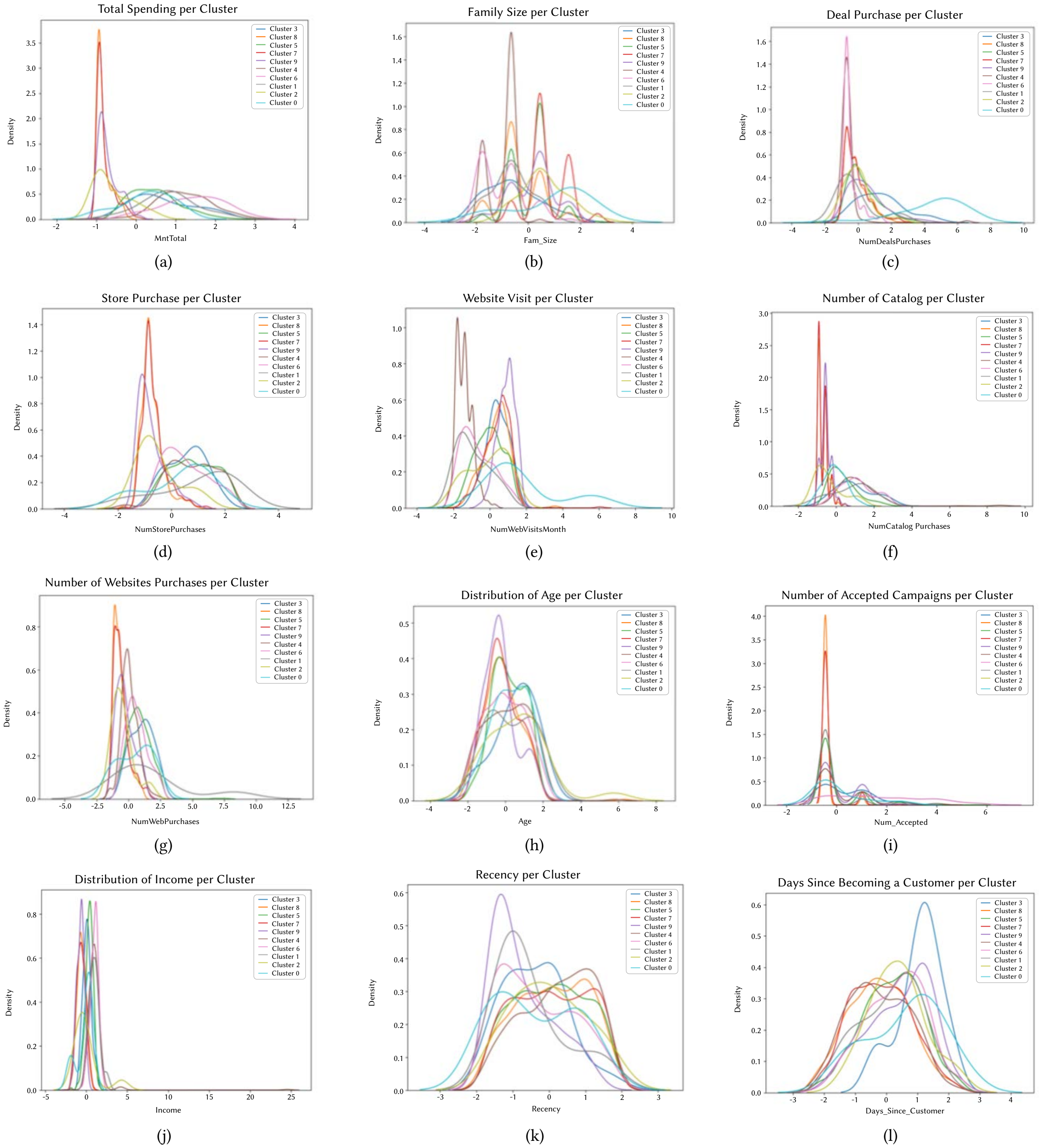
(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Fig. 12. DBSCAN clustering results across the Marketing Campaign dataset. Details are in section V.B.2, under DBSCAN.

size of approximately three members and are older, with an average age of around 56. They have a relatively low average income and a low total spending on products. These customers often make infrequent purchases, especially of meat and fish products, and are not very responsive to marketing campaigns. Cluster 2: These customers typically have medium incomes and moderate spending habits. They are mostly middle-aged, around 47 years old, with an average family size of about three members. Their purchase frequency is moderate, particularly for wines and sweets, and they show limited engagement with marketing efforts. They visit web stores quite often and are relatively consistent in their purchasing patterns. Cluster 3: This group has a slightly higher average age of 50 and consists of families with approximately three members. They have moderate incomes and spending levels, especially on wine, meat, and sweets. These customers show moderate responsiveness to marketing campaigns and have a balanced approach to both online and in-store purchases. Cluster 4: comprises smaller families, often single individuals or couples, with the lowest average income among the clusters. These customers are relatively young, around 41 years old, and have minimal spending, particularly on non-essential items like sweets and gold products. They show low engagement with marketing campaigns and visit online stores moderately. Cluster 5: With an average family size of nearly three, these customers have medium incomes and spending habits. They are generally middle-aged, around 50 years old, and exhibit moderate purchasing patterns, especially for wine and sweets. Their responsiveness to marketing campaigns is average, and they balance their shopping between online and physical stores. Cluster 6: This cluster consists of slightly larger families, around three members, with higher incomes and spending, particularly on wine and gold products. These customers are typically older, averaging 54 years in age, and show moderate engagement with marketing campaigns. They make frequent purchases both online and in-store, reflecting their active shopping behavior. Cluster 7: Customers in this cluster are older, averaging 58 years, with nearly three family members. They have high incomes and significant spending, particularly on wine, meat, and gold products. Their responsiveness to marketing campaigns is higher than



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

Fig. 13. GMM clustering results across the Marketing Campaign dataset. Details are in section V.B. 3, under GMM.

average, and they frequently shop both online and in-store, making them highly valuable customers. Cluster 8: These customers, averaging around 52 years old, have a slightly larger family size of about three members. Their income and spending levels are moderate, with a focus on wine and gold products. They show limited responsiveness to marketing campaigns but are consistent in their purchasing patterns, both online and in physical stores. Cluster 9: This cluster consists of families with approximately two to three members, averaging 45 years old. They have moderate incomes and spending, particularly on wine and gold products. These customers show a low engagement with marketing campaigns and have a balanced approach to online and in-store shopping. Cluster 10: Customers in this cluster are older, around 58 years, with moderate family sizes. They have high incomes and spend significantly, particularly on meat and fish products. Their engagement with marketing campaigns is above average, and they frequently shop both online and in physical stores, reflecting their active consumer behavior. Cluster 11: This group has high-income customers, typically around 54 years old, with smaller families. They exhibit high spending, especially on wine and gold products, and show moderate responsiveness to marketing campaigns. These customers visit online stores frequently and have a consistent purchasing pattern. Cluster 12: Customers in this cluster are older, averaging 57 years,

TABLE I. Comparison of Clustering Methods

Method	Dataset	Clusters	Calinski-Harabasz	Davies-Bouldin
K-means	Turkish Market Sales	6	–	1.60
K-means	Marketing Campaign	3	617.33	1.85
DBSCAN	Marketing Campaign	3	302.34	1.26
GMM	Marketing Campaign	10	184.92	2.29
SOM	Marketing Campaign	17	611.83	0.63

TABLE II. Summary of Clustering Results Across Algorithms (SP: Spending Patterns, FS: Family Size, PB: Purchasing Behaviors, AD: Age Distribution, MR: Marketing Responsiveness)

Attribute	K-means	DBSCAN	GMM	SOM
SP	High: Cluster 0 Medium: Cluster 1 Low: Cluster 2	High: Cluster 0 Medium: Cluster 2 Low: Cluster 1	High: Cluster 6 Medium: Clusters 1, 3, 5 Low: Clusters 0, 4, 7	High: Clusters 6, 7, 10, 12, 14 Medium: Clusters 2, 3, 5, 8, 9 Low: Clusters 1, 4
FS	Large: Cluster 1 Medium: Cluster 2 Small: Cluster 0	Large: Cluster 0 Medium: Cluster 2 Small: Outliers	Large: Cluster 0 Medium: Most clusters Small: Cluster 7	Large: Clusters 13, 15, 17 Medium: Most clusters Small: Clusters 4, 11, 14
PB	Online: N/A In-store: Clusters 0, 1 Catalog: Cluster 0	Online: Cluster 1 In-store: Cluster 2 Catalog: Cluster 1	Online: Cluster 1 In-store: Cluster 1 Catalog: Clusters 3, 6	Online: Clusters 2, 7, 12, 14 In-store: Clusters 3, 6, 10 Catalog: Varied
AD	Oldest: Cluster 2 (39) Middle: Cluster 1 (38) Youngest: Cluster 0 (32)	Oldest: Cluster 2 Middle: Cluster 1 Youngest: N/A	Oldest: Cluster 2 Middle: Most clusters Youngest: Clusters 7, 9	Oldest: Cluster 15 (60) Middle: Most clusters Youngest: Cluster 4 (41)
MR	High: Cluster 0 Medium: Cluster 1 Low: Cluster 2	High: Cluster 1 Medium: Outliers Low: Clusters 0, 2	High: Cluster 6 Medium: Clusters 3, 5, 8 Low: Clusters 0, 4, 7	High: Clusters 7, 12, 14 Medium: Clusters 3, 5, 8, 13 Low: Clusters 1, 4, 16

with smaller family sizes. They have high incomes and substantial spending, particularly on wine and gold products. They are highly responsive to marketing campaigns and exhibit frequent shopping behavior both online and in physical stores, making them highly valuable. Cluster 13: These customers are older, averaging around 58 years, with larger family sizes. They have moderate incomes and spending levels, particularly on wine and sweets. Their engagement with marketing campaigns is average, and they balance their shopping between online and physical stores. Cluster 14: This cluster consists of relatively younger customers, around 56 years old, with smaller family sizes. They have high incomes and significant spending, especially on wine, meat, and gold products. Their responsiveness to marketing campaigns is very high, and they frequently shop both online and in-store, making them among the most valuable customers. Cluster 15: Customers in this cluster are older, averaging 60 years, with larger family sizes. They have moderate incomes and spending levels, particularly on wine and sweets. Their engagement with marketing campaigns is average, and they balance their shopping between online and physical stores. Cluster 16: This group consists of middle-aged customers, around 49 years old, with medium family sizes. They have moderate incomes and spending levels, particularly on wine and sweets. Their responsiveness to marketing campaigns is low, and they show consistent purchasing patterns, both online and in physical stores. Cluster 17: These customers are older, around 53 years, with larger family sizes. They have moderate incomes and spending levels, particularly on wine and gold products. Their engagement with marketing campaigns is above average, and they balance their shopping between online and physical stores.

C. Comparative Analysis of Clusters

The application of four distinct clustering algorithms to the Marketing Campaign dataset revealed unique insights into customer segmentation:

- **K-means**: Provided a clear, income-based segmentation with three distinct clusters.
  - Uniquely identified a high-income, young (average 32 years) customer segment with high marketing responsiveness.
  - Revealed an inverse relationship between income and recency of purchases.
  - Emphasized the importance of purchasing behavior in defining customer segments.
  - Highlighted the correlation between high spending, frequent website visits, and marketing campaign acceptance.
- **DBSCAN**: Excelled in identifying outliers and non-spherical clusters.
  - Uncovered a distinct group of moderate-income, highly engaged customers (Cluster 2).
  - Uniquely categorized customers with variable spending patterns as outliers, potentially identifying niche market segments.
- **Gaussian Mixture Model (GMM)**: Provided the most granular segmentation with 10 distinct clusters.
  - Revealed subtle variations in customer behavior, particularly in the high-income segments.
  - Identified a unique cluster (Cluster 0) combining large family size, high deal responsiveness, and frequent website visits.
- **Self-Organizing Maps (SOM)**: Offered the most detailed age-based segmentation with 17 clusters.
  - Provided nuanced insights in age and consumer behavior (Cluster 6) as described in Section V.4.
  - Uniquely identified several high-value, older customer segments with distinct purchasing preferences.

The trade-off between interpretability and complexity is evident across our results. K-means provides straightforward, easily interpretable segments ideal for immediate business application, as evidenced by its clear three-cluster structure that retail managers can readily understand and act upon. DBSCAN maintains reasonable interpretability while adding the capability to identify outliers, offering a balance between simplicity and advanced clustering capabilities. In contrast, GMM (10 clusters) and SOM (17 clusters) offer significantly more granular segmentation but require additional analytical expertise to translate into actionable strategies. This increasing complexity allows for more nuanced understanding of customer behavior but may challenge practical implementation in retail environments where quick decision-making is essential.

Cross-algorithm comparisons revealed several key insights:

1. Income and age consistently emerged as primary factors in customer segmentation across all algorithms.
2. The inverse relationship between income and purchase recency was a common finding, particularly evident in K-means and DBSCAN results.
3. While K-means provided a broad overview with three clusters, GMM and SOM offered more granular insights, potentially useful for highly targeted marketing strategies.
4. DBSCAN's ability to identify outliers provided unique insights into niche customer groups that other algorithms might have overlooked.

Table I displays the results of various clustering methods applied to 2 different datasets, evaluating their performance based on the number of clusters identified, Calinski-Harabasz Index, and Davies-Bouldin score. The performance evaluation, using metrics such as the Calinski-Harabasz Index and the Davies-Bouldin Index, indicated that K-Means achieved the highest scores. Although GMM and SOM also yielded respectable scores, the highest CH score was achieved by K-Means, affirming its effectiveness for the datasets and objectives of this study.

Table II summarizes the clustering result across algorithms in a standardized notion. The composite attribute definitions used in the table are as follows:

- **SP: Spending Patterns**
  - **High**: Customers who spend the most money (high-value customers)
  - **Medium**: Customers with moderate spending levels
  - **Low**: Customers who spend the least (low-value customers)
- **FS: Family Size**
  - **Large**: Customers with big families
  - **Medium**: Customers with average-sized families (typically 2-3 members)
  - **Small**: Customers with small families or single-person households
- **PB: Purchasing Behaviors**
  - **Online**: Customers who prefer to shop through websites/online platforms
  - **In-store**: Customers who prefer to shop at physical store locations
  - **Catalog**: Customers who prefer to shop through catalogs (mail-order)
- **AD: Age Distribution**
  - **Oldest**: The older customer segments
  - **Middle**: Middle-aged customer segments
  - **Youngest**: The younger customer segments
- **MR: Marketing Responsiveness**
  - **High**: Customers who frequently respond to marketing campaigns, deals, and promotions
  - **Medium**: Customers with moderate response to marketing efforts
  - **Low**: Customers who rarely respond to marketing campaigns or promotions

VI. Conclusion and Future Work

This study conducted a comprehensive comparative analysis of four machine learning algorithms (K-means, DBSCAN, GMM, SOM) for customer segmentation in the Turkish retail market. Using two distinct datasets, a large-scale Turkish market sales dataset and a focused marketing campaign dataset, this research aimed to identify the most effective and actionable customer segmentation techniques for the unique characteristics of the Turkish retail sector.

K-means demonstrated the most robust performance, offering a balance between interpretability and statistical validity. DBSCAN showed strengths in identifying non-spherical clusters and handling outliers, while GMM and SOM provided more granular segmentation at the cost of increased complexity.

These findings have shown significant implications for Turkish retailers, enabling more targeted marketing strategies and improved customer relationship management. However, the study's limitations, including its focus on specific datasets, suggest caution in generalizing results.

An important consideration for retailers is the trade-off between model interpretability and complexity. Our findings demonstrate that while simpler algorithms like K-means offer highly interpretable results that can be readily implemented by marketing teams, more complex methods such as GMM and SOM provide deeper insights that may require specialized expertise to leverage effectively. Organizations must balance their need for sophisticated customer understanding against their capacity to interpret and act upon complex segmentation results.

Future work should explore the potential of deep learning techniques and hybrid models that combine traditional clustering approaches with neural networks, which could provide more sophisticated pattern recognition and potentially uncover complex, non-linear relationships in customer behavior data. These advanced approaches might include autoencoders for dimensionality reduction, deep clustering methods, or ensemble approaches that leverage the strengths of multiple algorithms.

Additionally, future research should explore the application of these algorithms across diverse retail sectors in Turkiye, investigate the long-term effectiveness of resulting marketing strategies, and examine how Turkish cultural norms, regional differences, and consumer behavior patterns influence segmentation strategies by analyzing how factors such as traditional shopping habits, family structures, and regional economic differences affect the interpretation and application of clustering results.

References

[1] P. Sharma, The ultimate guide to K-means clustering: Definition, methods and applications. Retrieved from https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/. Accessed on March, 2024.

[2] S. P. Nguyen, "Deep customer segmentation with applications to a Vietnamese supermarkets' data," *Soft Computtng*, vol. 25, no. 12, pp. 7785-7793, 2021.

[3] İ.Kabasakal, "Customer segmentation based on recency frequency monetary model: A case study in E-retailing," *Bilişim Teknolojileri Dergisi*, vol. 13, no. 1, pp. 47-56, 2020.

[4] G. Armstrong and P. Kotler, *Marketing: an introduction*, Pearson Educación, 2003.

[5] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1-35, 2021.

[6] T. Kansal, S. Bahuguna, V. Singh, and T. Choudhury, "Customer segmentation using K-means clustering," *2018 international conference on computational techniques, electronics and mechanical systems (CTEMS)*, pp. 135-139, IEEE, 2018.

[7] G. Mohit, *Customer Segmentation using Machine Learning applied to Banking Industry* (Doctoral dissertation, Hochschule Neu Ulm), 2023.

[8] D. Zakrzewska and J. Murlewski, "Clustering algorithms for bank customer segmentation," In *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, pp. 197-202, IEEE, 2005.

[9] C. Rungruang, P. Riyapan, A. Intarasit, K. Chuarkham, and J.Muangprathub, "RFM model customer segmentation based on hierarchical approach using FCA," *Expert Systems with Applications*, vol. 237, 121449, 2024.

[10] K. Tabianan, S. Velu, and V. Ravi, "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data," *Sustainability*, vol. 14, no. 12, 7243, 2022.

[11] A. Ashabi, S. B. Sahibuddin, and M. Salkhordeh Haghighi, "The systematic review of K-means clustering algorithm," In *Proceedings of the 2020 9th International Conference on Networks, Communication and Computing*, pp. 13-18, 2020.

[12] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters," in large spatial databases with noise. In *kdd*, Vol. 96, No. 34, pp. 226-231, 1996.

[13] V. Kachroo, "Customer segmentation and profiling for e-commerce using DBSCAN and fuzzy C-means," *Proceedings on Engineering*, vol. 5, no. 3, pp. 539-544, 2023.

[14] E. A. Laksana, and M. M. Fahrezi, "Customer segmentation and analysis based on Gaussian mixture model algorithm," In *Widyatama International Conference on Engineering 2024 (WICOENG 2024)*, pp. 67-75, Atlantis Press, 2024.

[15] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial-temporal data," *Data & knowledge engineering*, vol 60, no. 1, pp. 208-221, 2007.

[16] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, X. Xu, "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1-21, 2017.

[17] L. L. Scientific, "Data segmentation using mixture regression models with generalized Gaussian distribution and K-means," *Journal of Theoretical and Applied Information Technology*, vol. 103, no. 8, 2025.

[18] E. A. Laksana and M. M. Fahrezi, "Customer segmentation and analysis based on Gaussian mixture model algorithm," In *Widyatama International Conference on Engineering 2024 (WICOENG 2024)*, pp. 67-75, Atlantis Press, 2024.

[19] M. A. Camilleri and, M. A. Camilleri, *Market segmentation, targeting and positioning*, pp. 69-83, Springer International Publishing, 2018.

[20] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, 30, 2017.

[21] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake, Finite mixture models. *Annual review of statistics and its application*, vol. 6, no. 1, pp. 355-378, 2019.

[22] F. Saadi, B. Atmani, and F. Henni, "Improving retrieval performance of case based reasoning systems by fuzzy clustering," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 9, no. 1, pp. 84-91, 2024.

[23] Y. C. Liu, M. Liu, and X. L. Wang, "Application of Self- Organizing Maps in text clustering," *Applications of Self- Organizing Maps*, 205, 2012.

[24] D. Barman and N. Chowdhury, "A novel approach for the customer segmentation using clustering through self-organizing map," *International Journal of Business Analytics (IJBAN)*, vol. 6, no. 2, pp. 23-45, 2019.

[25] R. Vohra, J. Pahareeya, A. Hussain, F. Ghali, and A. Lui, "Using self organizing maps and K means clustering based on RFM model for customer segmentation in the online retail business," In *Intelligent Computing Methodologies: 16th International Conference, ICIC, Bari, Italy, Proceedings*, Part III 16, pp. 484-497, Springer International Publishing, 2020.

[26] S. Üstebey, İ. Yelmen, and M. Zontul, "Customer segmentation based on self-organizing maps: a case study on airline passengers," *Havacılık Ve Uzay Teknolojileri Dergisi*, 2020.

[27] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464-1480, 1990.

[28] I. Valova, G. Georgiev, N. Gueorguieva, and J. Olson, "Initialization issues in self-organizing maps," *Procedia Computer Science*, vol. 20, pp. 52-57, 2013.

[29] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society: series B (methodological)*, vol. 39, no. 1, pp. 1-22, 1977.

[30] K. Kumar Sharma, A. Seal, A. Yazidi, and O. Krejcar, "S- divergence-based internal clustering validation index," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4, pp. 127-139, 2023.

[31] O. Colakoglu, 10 Million Rows Turkish Market Sales Dataset (MSSQL). Retrieved from https://www.kaggle.com/datasets/omercolakoglu/10million-rows-turkish-market-sales-dataset. Accessed: March 12, 2024.

[32] R. Saldanha, Marketing Campaign. Retrieved from https://www.kaggle.com/datasets/rodsaldanha/ arketing-campaign. Accessed on March, 2024.

Nur Diyabi

She received a B.Sc. degree in computer engineering with a minor in strategic public relations management from Bahcesehir University in Türkiye in 2022 and M.Sc. degree in big data analytics from Bahcesehir University in Türkiye, in 2024. She is currently working as a Security Engineer at Paramount Computer Systems in UAE.

Duygu Çakır

She was born in Istanbul, Turkiye. She received her BSc, MSc, and PhD degrees in computer engineering department in Bahcesehir University (BAU), Turkiye. She also finished her under-grad level double major in mathematics and computer sciences. From 2007 to 2012, she was a Research Assistant in computer and software engineering departments respectively. Since 2012, she has been working in the department of Software Engineering. She worked in government projects, managed many software and artificial intelligence related projects, and has experience in introductory and advanced programming, data structures, computer graphics, and data science courses in Istanbul as a full time Assistant Professor, in Berlin-Germany and Jelgava-Latvia as a visiting professor. Her research interests include facial action unit and facial expression analysis as well as automated machine learning in computer vision and she holds a national patent on eye tracking in mobile devices, another national patent (pending) on generating virtual agents by converting the input voice directly to a non-existing synthetic face.

Ömer Melih Gül

He received BSc., MSc., and PhD. degrees from the Department of Electrical and Electronics Engineering at Middle East Technical University (METU), Ankara, Türkiye, in 2012, 2014, and 2020, respectively by also working as a research assistant at the same department. His research interests include AI/machine learning applications, wireless security, networking, scheduling, IoT, UAV, robotics, and blockchain. He has co-authored over 50 papers and 4 book chapters. He was awarded third place in the 2019 Lance Stafford Larson Outstanding Student Paper Award by the IEEE Computer Society. He was also awarded third place in the poster competition at 2021 IEEE Rising Stars Global Conference. In 2022, he worked as a postdoctoral fellow at School of Electrical Engineering and Computer Science at University of Ottawa, Canada. He is a recipient of the best paper award at 48th Wireless World Research Forum (WWRF) in 2022. In 2023, he worked as an Assistant Professor in the Department of Computer Engineering at Bahcesehir University, Istanbul, Türkiye, where he supervised 4 MSc theses and co- supervised 1 thesis. Since March 2024, he has been working as an associate professor in the Informatics Institute at Istanbul Technical University (ITU), Istanbul, Türkiye, where he is supervising 1 PhD and 4 MSc students. He serves as an Editor in IEEE Open Journal of Computer Society, (Elsevier) Sustainable Computing: Informatics and Systems, (Springer) Telecommunication Systems, Wireless Networks, Cluster Computing and also the International Journal of Interactive Multimedia and Artificial Intelligence. As cochair, he organized CIEAI workshop at IEEE Fog and Mobile Edge Computing (FMEC) 2023 in Estonia. Moreover, he became the Publicity Chair in the IEEE iThings 2024. He organized two editions of EAI International Conference on Robotic Sensor Networks (ROSENET 2023, ROSENET 2024) as general chair.

Tevfik Aytekin

He is an Associate Professor in the Department Computer Engineering at Bahçeşehir University, in İstanbul, Turkey. He received his Ph.D. in Cognitive Science from Middle East Technical University, M.Sc. in Computer Science from Hacettepe University, and B.Sc. in Computer Science from Bilkent University. His current research interests include data mining, machine learning, and recommender systems. In addition to his academic work, he actively consults on AI/ML projects across various industries.

Seifedine Kadry

He has a bachelor's degree in 1999 from Lebanese University, MS degree in 2002 from Reims University (France) and EPFL (Lausanne), PhD in 2007 from Blaise Pascal University (France), HDR degree in 2017 from Rouen University. He is a Full Professor of Data Science at Lebanese American University, Lebanon. At present, his research focuses on Data Science, education using technology, system prognostics, stochastic systems, and applied mathematics. He is an ABET program evaluator for Computing and an ABET program evaluator for Engineering Tech. He is a Fellow of IET, Fellow of IETE, and Fellow of IACSIT. He was a distinguished speaker of the IEEE Computer Society.

unir
LA UNIVERSIDAD
EN INTERNET

Rectorado
Avenida de la Paz, 137
26006 Logroño (La Rioja)
t (+34) 941 21 02 11

www.unir.net