International Journal of Interactive Multimedia and Artificial Intelligence

June 2024, Vol. VIII, Number 6
ISSN: 1989-1660

unir LA UNIVERSIDAD EN INTERNET

> *"Artificial intelligence has the potential to vastly improve the quality of life for billions of people all over the world. If we get it right, AI could be the best thing ever for humanity."*
> *Demis Hassabis, Co-founder and CEO of DeepMind*

INTERNATIONAL JOURNAL OF INTERACTIVE MULTIMEDIA AND ARTIFICIAL INTELLIGENCE

ISSN: 1989-1660 –VOL. 8, NUMBER 6

EDITORIAL TEAM

Editor's Note

The International Journal of Interactive Multimedia and Artificial Intelligence – IJIMAI (ISSN 1989-1660) provides an interdisciplinary forum in which scientists and professionals can share their research results and report new advances in Artificial Intelligence (AI) tools or tools that use AI with interactive multimedia techniques.

This regular issue consists of 16 articles that use artificial intelligence or computational systems to come up with new solutions and solve problems more effectively. The issue showcases the use of Artificial Intelligence or computational systems that contribute to new knowledge with innovative applications. In this issue you can find different articles on game theory, models for collaborative filtering, text classification, fake news detection system, identification system, semi eager classifier, longitudinal segmented analysis, etc.

The issue begins with a review of the main studies about the Game Theory in Quantum Computers, by Raquel Pérez-Antón et al., including operational requirements and implementation details. In addition, the article describes various quantum games, their design strategy, and the used supporting tools. They also present the still open debate linked to the interpretation of the transformations of classical algorithms in fundamental game theory to their quantum version, with special attention to the Nash equilibrium.

The next article of this volume is of Jesús Bobadilla et al. The title is "Comprehensive Evaluation of Matrix Factorization Models for Collaborative Filtering Recommender Systems". This article tested six representative matrix factorization models, using four collaborative filtering datasets. Experiments have tested a variety of accuracy and beyond accuracy quality measures, including prediction, recommendation of ordered and unordered lists, novelty, and diversity. Results show each convenient matrix factorization model attending to their simplicity, the required prediction quality, the necessary recommendation quality, the desired recommendation novelty and diversity, the need to explain recommendations, the adequacy of assigning semantic interpretations to hidden factors, the advisability of recommending to groups of users, and the need to obtain reliability values.

Then we find the work of Raúl A. del Águila Escobar et al. with the title "OBOE: an Explainable Text Classification Framework". This article presents a text classification framework called OBOE (explanatiOns Based On concEpts), in which such ingredients play an active role to open the black-box. OBOE defines different components whose implementation can be customized and, thus, explanations are adapted to specific contexts. They also provide a tailored implementation to show the customization capability of OBOE. Additionally, they performed (a) a validation of the implemented framework to evaluate the performance using different corpora and (b) a user-based evaluation of the explanations provided by OBOE.

The following research presents: An Efficient Fake News Detection System Using Contextualized Embeddings and Recurrent Neural Network, built by Junaid Ali Reshi and Rashid Ali. This study focuses on efficient detection of fake news on social media, through a natural language processing based approach, using deep learning. For the detection of fake news, textual data have been analyzed in unidirectional way using sequential neural networks, or in bi-directional way using transformer architectures like Bidirectional Encoder Representations from Transformers (BERT). This article proposes Contextualized Fake News Detection System (ConFaDe) - a deep learning based fake news detection system that utilizes contextual embeddings generated from a transformer-based model. The model uses masked language modelling and replaced token detection in its pre-training to capture contextual and semantic information in the text.

Miguel García García et al. present: Graffiti Identification System Using Low-Cost Sensors. This article introduces the possibility of studying graffiti using a colorimeter developed with Arduino hardware technology according to the Do It Yourself (DIY) philosophy. Through the obtained Red Green Blue (RGB) data it is intended to study and compare the information extracted from each of the graffiti present on different walls. The same color can be found in different parts of a single graffiti, but also in other graffiti that could a priori be of different authorship. Nevertheless, graffiti may be related, and it may be possible to group graffiti artists and "gangs" that work together.

We then find the article titled "PeopleNet: A Novel People Counting Framework for Head-Mounted Moving Camera Videos" by Ankit Tomar et al. This study proposes a transfer learning-based PeopleNet model to tackle people counting problem. For this, they have made some significant changes to the standard VGG16 model, by disabling top convolutional blocks and replacing its standard fully connected layers with some new fully connected and dense layers. The strong transfer learning capability of the VGG16 network yields in-depth insights of the PeopleNet into the good quality of density maps resulting in highly accurate crowd estimation.

The next research of Sami Dhahbi et al. is titled "Lightweight Real-Time Recurrent Models for Speech Enhancement and Automatic Speech Recognition". This study proposes a lightweight hourglass-shaped model for speech enhancement (SE) and automatic speech recognition (ASR). Simple recurrent units (SRU) with skip connections are implemented where attention gates are added to the skip connections, highlighting the important features and spectral regions. The model operates without relying on future information that is well-suited for real-time processing. Combined acoustic features and two training objectives are estimated. Experimental evaluations using the short time speech intelligibility (STOI), perceptual evaluation of speech quality (PESQ), and word error rates (WERs) indicate better intelligibility, perceptual quality, and word recognition rates.

A. Suruliandi et al. present the article "Drug Target Interaction Prediction Using Machine Learning Techniques – A Review". This article aims to explore Machine Learning (ML) techniques better for Drug Target Interaction (DTI) prediction and boost future research. Qualitative and quantitative analyses of ML techniques show that several have been applied to predict DTIs, employing a range of classifiers. Though DTI prediction improves with negative Drug Target Pairs (DTP), the lack of true negative DTPs has led to the use a particular dataset of drugs and targets. Using dynamic DTPs improves DTI prediction. Little attention has so far been paid to developing a new classifier for DTI classification, and there is, unquestionably, a need for better ones.

M. Akshay Kumaar et al. present the article titled "Brain Tumor Classification Using a Pre-Trained Auxiliary Classifying Style-Based Generative Adversarial Network". This research proposes a novel approach that uses a style-based generative adversarial network for conditional synthesis and auxiliary classification of brain tumors by pre-training. The discriminator of the pre-trained GAN is fine-tuned with extensive data augmentation techniques to improve the classification accuracy when the training data is small. The proposed method was validated with an open-source Magnetic Resonance Imaging (MRI) dataset which consists of three types of tumors - Glioma, Meningioma, and Pituitary. The proposed system achieved

DOI: 10.9781/ijimai.2024.05.002

99.51% test accuracy, 99.52% precision score, and 99.50% recall score, outperforming other approaches. Since the framework can be made adaptive using transfer learning, this method also benefits new and small datasets of similar distributions.

In the article titled "KoopaML: A Graphical Platform for Building Machine Learning Pipelines Adapted to Health Professionals", F.J. García-Peñalvo et al. present a platform to assist non-expert users in defining ML pipelines in the health domain. The system's design focuses on providing an educational experience to understand how ML algorithms work and how to interpret their outcomes, and on fostering a flexible architecture to allow the evolution of the available components, algorithms, and heuristics.

The title of the next article is "GRASE: Granulometry Analysis With Semi Eager Classifier to Detect Malware". Mahendra Deore et al. propose a malware classification using a visualization methodology wherein the disassembled malware code is transformed into grey images. They present the efficacy of granulometry texture analysis technique for improving malware classification. Furthermore, a Semi Eager (SemiE) classifier, which is a combination of eager learning and lazy learning technique, is used to get robust classification of malware families. The outcome of the experiment is promising since the proposed technique requires less training time to learn the semantics of higher-level malicious behaviours. Identifying the malware (testing phase) is also done faster. A benchmark database like malimg and Microsoft Malware Classification challenge (BIG-2015) has been utilized to analyse the performance of the system. An overall average classification accuracy of 99.03 and 99.11% is achieved, respectively.

In the work titled "Chatbot-Based Learning Platform for SQL Training", Antonio Balderas et al. propose a chatbot-based learning platform to assist students in learning SQL. A case study has been conducted to evaluate the proposal, with undergraduate computer engineering students using the learning platform to perform SQL queries while being assisted by the chatbot. The results show evidence that students who used the chatbot performed better on the final SQL exam than those who did not. In addition, the research shows positive evidence of the benefits of using such learning platforms to support SQL teaching and learning for both students and lecturers: students use a platform that helps them self-regulate their learning process, while lecturers get interesting metrics on student performance.

Juan Antonio Caballero-Hernández et al. show the work: Supporting Skill Assessment in Learning Experiences Based on Serious Games Through Process Mining Techniques. They propose an automated method to analyse students' interactions and assess their skills in learning experiences based on serious games. The method takes into account not only the final model obtained by the student, but also the process followed to obtain it, extracted from game logs. The assessment method groups students according to their in-game errors and in-game outcomes. Then, the models for the most and the least successful students are discovered using process mining techniques. Similarities in their behaviour are analysed through conformance checking techniques to compare all the students with the most successful ones. Finally, the similarities found are quantified to build a classification of the students' assessments. They have employed this method with Computer Science students playing a serious game to solve design problems in a course on databases. The findings show that process mining techniques can palliate the limitations of skill assessment methods in game-based learning experiences.

The next article of Andreas Hinderks et al. is titled "Requirements for User Experience Management - A Tertiary Study". The work explains that when applied to User eXperience (UX), user experience management consists of a UX goal, a UX strategy, and UX resources. The authors conducted a tertiary study and examined the current state of existing literature regarding possible requirements. They want to figure out what requirements can be derived from the literature reviews with the focus on UX and agile development. In total, they were able to identify and analyse 16 studies. After analysing the studies in detail, they identified 13 different requirements for UX management. The most frequently mentioned requirements were prototypes and UX/usability evaluation. Communication between UX professionals and developers was identified as a major improvement in the software development process.

The following research is titled "Longitudinal Segmented Analysis of Internet Usage and Well-Being Among Older Adults". The authors Alejandro Cervantes et al. analyze a sample of 2,314 individuals, aged 50 years and older, that participated in the English Longitudinal Study of Aging. Participants were clustered according to drivers of psychological well-being using Self-Organizing Maps. The resulting groups were subsequently studied separately using generalized estimating equations fitted on 2-year lagged repeated measures using three scales to capture the dimensions of well-being and Markov models. The clustering analysis suggested the existence of four different groups of participants. Statistical models found differences in the connection between internet use and psychological well-being depending on the group. The Markov models showed a clear association between internet use and the potential for transition among groups of the population characterized, among other things, by higher levels of psychological well-being.

This issue finishes with the research titled "Modulating the Gameplay Challenge Through Simple Visual Computing Elements: A Cube Puzzle Case Study" by Jose Ribelles et al. In this work, a modulating mechanism based on visual computing is explored. The main hypothesis is that simple visual modifications of some elements in the game can have a significant impact on the game experience. This concept, which is essentially unexplored in the literature, has been experimentally tested with a web-based cube puzzle game where participants played either the original game or the visually modified game. The analysis is based on players' behavior, performance, and replies to a questionnaire upon game completion. The results provide evidence on the effectiveness of visual computing on gameplay modulation. The findings are relevant to game researchers and developers because they highlight how a core gameplay can be easily modified with relatively simple ingredients, at least for some game genres. Interestingly, the insights gained from this study also open the door to automate the game adaptation based on observed player's interaction.

Dr. Carlos Enrique Montenegro-Marin
Associate Editor
Universidad Distrital Francisco José de Caldas, Bogotá, Colombia

TABLE OF CONTENTS

EDITOR'S NOTE 3

THE GAME THEORY IN QUANTUM COMPUTERS: A REVIEW 6

COMPREHENSIVE EVALUATION OF MATRIX FACTORIZATION MODELS FOR COLLABORATIVE FILTERING RECOMMENDER SYSTEMS 15

OBOE: AN EXPLAINABLE TEXT CLASSIFICATION FRAMEWORK 24

AN EFFICIENT FAKE NEWS DETECTION SYSTEM USING CONTEXTUALIZED EMBEDDINGS AND RECURRENT NEURAL NETWORK 38

GRAFFITI IDENTIFICATION SYSTEM USING LOW-COST SENSORS 51

PEOPLENET: A NOVEL PEOPLE COUNTING FRAMEWORK FOR HEAD-MOUNTED MOVING CAMERA VIDEOS 61

LIGHTWEIGHT REAL-TIME RECURRENT MODELS FOR SPEECH ENHANCEMENT AND AUTOMATIC SPEECH RECOGNITION 74

DRUG TARGET INTERACTION PREDICTION USING MACHINE LEARNING TECHNIQUES – A REVIEW 86

BRAIN TUMOR CLASSIFICATION USING A PRE-TRAINED AUXILIARY CLASSIFYING STYLE-BASED GENERATIVE ADVERSARIAL NETWORK 101

KOOPAML: A GRAPHICAL PLATFORM FOR BUILDING MACHINE LEARNING PIPELINES ADAPTED TO HEALTH PROFESSIONALS 112

GRASE: GRANULOMETRY ANALYSIS WITH SEMI EAGER CLASSIFIER TO DETECT MALWARE 120

CHATBOT-BASED LEARNING PLATFORM FOR SQL TRAINING 135

SUPPORTING SKILL ASSESSMENT IN LEARNING EXPERIENCES BASED ON SERIOUS GAMES THROUGH PROCESS MINING TECHNIQUES 146

REQUIREMENTS FOR USER EXPERIENCE MANAGEMENT - A TERTIARY STUDY 160

LONGITUDINAL SEGMENTED ANALYSIS OF INTERNET USAGE AND WELL-BEING AMONG OLDER ADULTS 168

MODULATING THE GAMEPLAY CHALLENGE THROUGH SIMPLE VISUAL COMPUTING ELEMENTS: A CUBE PUZZLE CASE STUDY 177

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.*

The Game Theory in Quantum Computers: A Review

Raquel Pérez-Antón, José Ignacio López Sánchez, Alberto Corbi *

Universidad Internacional de La Rioja, Avd. de la Paz, 137 26006 Logroño, La Rioja (Spain)

* Corresponding author: raquel.perez527@comunidadunir.net (R. Pérez-Antón), joseignacio.lopez@unir.net (J. I. López Sánchez), alberto.corbi@unir.net (A. Corbi Bellot).

Received 6 July 2021 | Accepted 26 July 2023 | Early Access 19 September 2023

Abstract

Game theory has been studied extensively in recent centuries as a set of formal mathematical strategies for optimal decision making. This discipline improved its efficiency with the arrival, in the 20th century, of digital computer science. However, the computational limitations related to exponential time type problems in digital processors, triggered the search for more efficient alternatives. One of these choices is quantum computing. Certainly, quantum processors seem to be able to solve some of these complex problems, at least in theory. For this reason, in recent times, many research works have emerged related to the field of quantum game theory. In this paper we review the main studies about the subject, including operational requirements and implementation details. In addition, we describe various quantum games, their design strategy, and the used supporting tools. We also present the still open debate linked to the interpretation of the transformations of classical algorithms in fundamental game theory to their quantum version, with special attention to the Nash equilibrium.

Keywords

Nash Equilibrium, Polynomial Time Quantum Problems, Quantum Computing, Quantum Game Strategies, Quantum Game Theory.

DOI: 10.9781/ijimai.2023.09.001

I. Introduction

The application areas of quantum computing cover disciplines such as chemistry, physics, artificial intelligence and data mining, among others. Currently, the most relevant studies are related to the field of cybersecurity, with the aim of verifying if classic cryptographic systems are robust enough to face quantum computer attacks. However, other applications in economics, finance, geopolitics, psychology, and even human behaviour, are drawing the attention of the research community. In these disciplines, it is common to use classic algorithms for decision-making and payment strategies, which are intrinsically associated with game theory.

Game theory, or *interactive decision theory*, is considered the formal technique for decision making. Although previously, some authors began their research around its formal outline, it was not until 1944 when Von Neumann and Morgenstern [1] proposed a mathematical structure based on set theory, prepositional logic, matrix algebra, linear geometry and group theory. The incipient study on the transformation of classical to quantum algorithms, and their potential entanglement in multiple strategies, has also been involved in the advancement of quantum game theory. However, the difficulty of these transformations sometimes resides in the design restrictions of quantum circuits, defined by DiVincenzo in his article [2]. These restrictions not only force the initialization of the game scenarios in a different way compared to their classic version, but they also affect the way the game is played as time evolves. During the game, there are player-related movements that are difficult to reproduce. In the real world, any given environment or *scenario* is continuously changing as the pursued strategies evolve.

In this work, we review some of the main studies related to quantum game theory algorithms and the different interpretations in its transformation from its classical orchestration. In addition, we address the main techniques and procedural background used in this area of knowledge. In this context, the present work complements other review efforts on the subject, such as that carried out by Guo *et al.* [3].

A. The Game Theory and Computer Science

Game theory, or the discipline associated to the *search for optimal decision strategies to maximize profits*, has shaped other areas of knowledge, mainly Mathematics and military strategy. Originally, games were adapted to the latter, since choosing an appropriate strategy provides an advantage over other opponents. The origins of games such as chess in the 18th century in Prussia, served as a means of teaching future army officers concepts assimilable to military tactics [4]. Throughout the centuries, we find examples of these tools such as the Scytale used during the war between Athens and Sparta 431 BC, the Caesar cipher used 100 BC, or Enigma used in WWII by the German army. However, the best example of a military communication instrument is, perhaps, Arpanet, created by the United States Department of Defense in the middle of the Cold War with the Soviet Union, which forged the foundations of the current Internet. That is why, through the centuries, we observe that the application of game theory has been supported with the cutting-edge technological tools and knowhow of each era.

Please cite this article as:
R. Pérez-Antón, J. I. López Sánchez, A. Corbi Bellot, "The Game Theory in Quantum Computers: A Review", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 6, pp. 6-14, 2024, http://dx.doi.org/10.9781/ijimai.2023.09.001

In the 20th century, the first programmable processors were also applied to game theory given the increase in calculation speed. For instance, C. Shannon proposed chess as a testing ground for the development of artificial intelligence in 1950 [5]. In 1996, G. Kasparov, the world's best chess player at that time, was defeated by the Deep Blue supercomputer. This piece of technology, created by IBM, was capable of forecasting 100 million plays per second.

Nevertheless, it is widely known that classical Turing Machines (TM) suffer from computational limitations [6],[7]. Currently, TM is accepted as the correct formalization of the algorithm concept. An algorithm is a sequence of executable logical steps that allow solving a problem. It must also meet two properties: have a finite description and be made up of discrete operations that can be mechanically executed [8]. Briefly, we will define TM as a theoretical concept with a series of deterministic mechanical execution steps on an infinite tape on both sides with read/write head. These execution steps are defined and collect all the computational finite automaton processing of the input string. The TM continues to perform execution steps until it reads a symbol for which no action is defined. Whenever a TM accepts an input string the machine stops. However, a string is not accepted if it is stopped in a non-accepting state or by infinite looping (it never stops). In the latter case we cannot know if TM rejects the chain. A TM computes *f*, where *f* is a decision problem (not a function problem), in time $T(n)$, where $n \geq 0$ finite if its computation in each input string x needs at most $T(|x|)$ steps.

That is why this hypothetical machine serves as a measurement tool to determine the limitations and complexities that can be addressed by classical computers. There are variations to the original TM for solving problems such as the multitrack variation, with bounded memory register, multi-tape, non-deterministic, *or the quantum TM.* All of them measure the computational limitation of an algorithm, in time and resources.

B. Decision Problems in Classical Vs. Quantum Computers

Decision problems belong to the formal mathematical realm of game theory. That is why any algorithm with the aim of solving this type of problem can be said to belong to game theory or interactive decision theory. As stated above, the execution of an algorithm has computational limitations, therefore, the algorithms for solving games or making decisions will also have them.

The theory of complexities or computational limitations were introduced in several articles by Hartmains and Stearns in 1965 [9]. Computational complexity is responsible for analysing the resources, time, and memory to solve a problem. The main objective of the computational complexity theory is to identify the processing limits. That is why the analytical comparison between limitations of classical and quantum computers is essential. In a classical computer, both low-level circuits (hardware) and high-level programs (software) act under a structure based on algorithms that solve problems iteratively (step by step). Finding the algorithm that efficiently solves a problem is synonymous with finding the minimum consumption of time and resources. That is why the scientific community has determined that algorithms that are solved in polynomial time are efficient and in exponential time are intractable. Although some intractable decision problems, it seems that they are possible to be solved by quantum processors. The mathematical notation for representing spatial and temporal complexity when $n \to \infty$ in the worst case is defined as O (big O-notation). The scale of complexities is defined as (1):

$$\begin{aligned} O(1) \subset O(lg\, lg\, n) \subset O(ln(n) \; \subset O\,(lg^{a>1} n) \subset O(\sqrt{n}) \\ \subset O(n) \; \subset O\,(n\, lg\, n) \; \subset O(n^2) \\ \subset \cdots \; \subset O(n^{a>2}) \subset O(2^n) \; \subset O(n!) \\ \subset O\,(n^n) \end{aligned} \qquad (1)$$

TABLE I. Response Time for Two Values of the Size and Complexities n^3 and 2^n for A Step Value = 0.1 Millisecond

Complexity	$n = 32$	$n = 64$
n^3	3 secs	26 secs
2^n	5 days	$25 * 10^6$ years

If $O(g)$ is the asymptotic upper bound of the complexity of any algorithm and c is a positive constant of factors external to the algorithm, such as the machine to be executed. We have, as it is shown in Fig. 1, the polynomial time is the temporal complexity function $f(x)$, where x is the number of algorithm step value instructions. Therefore, the function f belongs to the complexity class of g ($f \in O(g)$ if there exists a c and an x_0 such that for all $x \geq x_0$ we have $|f(x) \leq c\,|\,g(x)|$ [10].

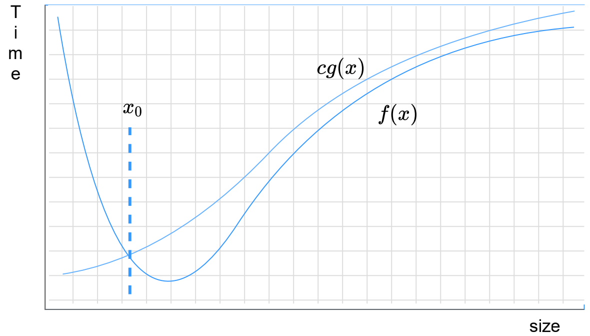


Fig. 1. Generic graphical representation of the comparison of functions in the calculation of complexities of any algorithm [11], including the problems generated by game theory.

Therefore, thanks to set theory, the characteristics common to all those decision problems with resolution in polynomial time and common characteristics can be grouped. The grouping of this type of mathematical problems according to their computational complexity or computational limitational are called *classes*, and their interrelation can be clearly seen in Fig. 2. In this way, we define some of the classes of fundamental decision-type problems based on their level of computational complexity:

P problems can be solved in polynomial time (linear, quadratic, cubic, etc.) in a deterministic MT. That is, the total time required by a processor to solve a problem that is bounded by a polynomial as a function of the size of the input and the number of configurations of its output.

NP problems are defined in a polynomial time. NP problems can be found in graph theory such as isomorphism or Hamiltonian paths.

PP The problems are solved in probabilistic polynomial time measured in a probabilistic MT. That is, the result obtained has an error with a probability of less than ½ for all cases. An example of an algorithm solved under probabilistic polynomial time is the Solovay-Strassen test [12].

BQP are problems that can be solved in polynomial time in a *quantum* TM. That is, the total time required by a quantum processor to solve a problem based on the size of the input (number of qubits), the number of configurations of its output and a maximum 1/3 probability of error for all instances and, therefore a success of 2/3. For example, we can find Shor's integer factorization algorithm in this class of problems [13].

BPP problems can be solved in probabilistic polynomial time measured in a probabilistic TM. That is, the result obtained has an error with a probability 1/3 and a success of 2/3. This type of problem is opposed to the Knapsack problem (KP) of combinatorial organization [14] solved by BQP, and all its elements must be included in the proposal for its resolution.

There are more kinds of decision problems such as NP-Complete, EXPTIME, L or NL, although we will not address them in this work. The class PH was determined by Larry Stockmeyer [15], and it unifies all the classes of hierarchical polynomial complexity. The potential of quantum computing is that, according to Aaronson [16], there is some evidence that BQP is not contained in PH. This entails that we could be approaching the resolution of exponential problems, intractable until now by classical computing.

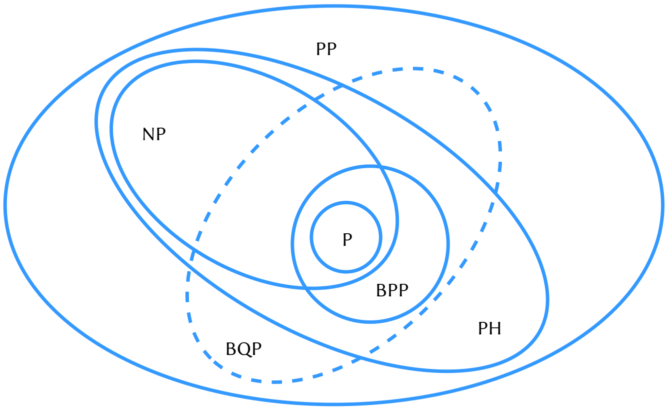


Fig. 2. Diagram of relationships between the different classes of complexities [17]. The BQP class can solve some problems in polynomial time that are not contained in PH.

C. Decision Strategies in Game Theory

Game theory has its own defined and structured decision problems according to its peculiarities. The types of games, their decision strategies, the format of the scenario and the number of participants are some of the characteristics of classical modelling. These strategy models correspond to a specific mathematical problem included in polynomial or exponential classes. Therefore, the correct initial definition of the classical strategy will be transformed into a class of computational complexity or limitation defined by both the quantum and the classical realms. Some of these strategies and payoff models are essential for understanding the transformation from classical to quantum algorithms.

Cooperatives and Non-Cooperatives strategies are characterized by the realization of alliances between the players with the objective of intensifying the maximum common benefit. These types of strategies can be found applied in common population resources such as a recycling plant, desalination plants, fire brigades, etc. Where the contribution of each player amplifies the benefit obtained by all participants. These cooperative strategies can also be used in areas such as politics, geopolitics, economics, armed conflicts, national and international markets. On the contrary, non-cooperative strategies are defined as those used by each player with the objective of satisfying individual benefit.

Sum 0 is closely linked to the interdependence between payments. Taking as the absolute factor payment to be distributed among all the players, a 0-sum game is understood to be one that the benefit of one player affects the losses of all the others. In other words, what a player has won necessarily comes from what another player or players have lost. This concept is very widespread in the financial world since the pie to be shared is finite and whenever an investor in the stock market gains the profit is associated with the loss of another individual. The opposite concept is *non-zero-sum games*. They are defined as those in which the cooperation between the participants of the game generates an equal and common benefit or loss.

Nash equilibrium was theorized by John. F Nash, who was later awarded the Nobel Prize in Economics in 1994 for his equilibrium analysis in non-cooperative game theory developed in 1950 [18]. The Nash equilibrium in game theory is becoming the most prominent unifying theory in the social sciences as indicated in his article [19]. Nash introduced the concept based on the relationship between the strategic equilibrium of the players and their maximum profit [20], which reads: “any n-tuple of strategies, one for each player, may be regarded as a point in the product space obtained by multiplying the n strategy spaces of the players. One such n-tuple counters another if the strategy of each player in the countering n-tuple yields the highest obtainable expectation for its player against the $n - 1$ strategies of the other players in the countered n-tuple”. A self-countering n-tuple is called an *equilibrium point*. In this strategy, it is assumed that all players know each other's strategies and do not cooperate with each other. In addition, the best strategy of a player is not synonymous with maximum payout but with less loss or 0 losses. Other essential concepts for understanding the Nash equilibrium are pure or mixed strategies. The pure strategies are those that each player chooses with probability 1, as an example we have the game of *rock, paper, and scissors*. In this game, each player selects his strategy based on a single payment. However, if we assign to each pure strategy a probability on the payout, we will be defining the mixed strategies. Mixed strategies are a generalization of pure strategies, therefore, in each one we can find a pure one. Nash showed that any finite rectangular game has at least one Nash equilibrium in mixed strategies [20].

Pareto optimal is one of the fundamental theories of welfare economics and was introduced by Vilfredo Pareto in 1896 [21]. And it is currently applied in different areas such as operations research, decision making, optimization with multiple objectives or cost-benefit analysis. It consists in that, given an initial allocation of earnings among a set of players, a change towards a new allocation that at least improves the situation of one individual without making the situation of the others worse is called *improvement*. An allowance is defined as *Pareto optimal* when no further improvements can be achieved. Therefore, it is no longer possible to benefit more individuals in a system without harming others. The *Pareto frontier* is identified with the function $f(x)$, where when expanding its domain, the gain of an individual is a consequence of the decrease of another participant. We formally define the concept as: let P be a multi-objective optimization problem, then a solution P_i is the Pareto optimal when there is no other solution P_j such that it improves on one objective without worsening at least one of the others.

Stochastics games were originally devised by Shapley [22]. They consist of achieving different states of the game system in time. That is, over time the choice of strategies of the players are conditioned to the current state or set of variables. Fundamental examples of these games are dice-based. They could be treated as games where the chance or different variables change the players' choice of strategies and payout over time. This game model has its application in market economies, such as the stock market. Furthermore, stochastic games can be approached from different perspectives, such as finite or unlimited in time, with partial information to the players, non-cooperative or 0-sum, among others.

D. Characteristics of Quantum Computers

Quantum computing is a different computing paradigm from digital computing. The internal logic architecture of a classical computer works with electrical pulses that are translated into high voltage 1 and low voltage 0, with deterministic input and output. However, in a quantum computer the conceptual change of its internal structure is based on quantum mechanics, not deterministic. Therefore, it is essential to review some of these concepts to get closer to this new computational model.

The main characteristic of quantum processors lies in the ability to manipulate quantum bits, known as *qubits*. Qubits can be represented by subatomic particles like electrons or photons with the intrinsic characteristics of quantum mechanics. That is why quantum computers use the properties of entanglement, superposition, and parallelism to optimize computational processing. The concept of qubit is not associated with a specific physical system, and they are described as a unit module vector in a complex two-dimensional vector space. The two basic states are are $|0\rangle$ and $|1\rangle$, but qubits can also be found in a state of superposition [23].

The superposition is associated with each physical system, where there is a Hilbert space (H) known as the state space of the system. The system is completely described by its state vector (represented in (2)), which is a unit vector in that state space. The states of the qubit represent a vector of states in a vector of states in the Hilbert 2-D space (H^2) with an orthonormal base.

$$|0\rangle = \begin{pmatrix}1\\0\end{pmatrix} \qquad |1\rangle = \begin{pmatrix}0\\1\end{pmatrix} \tag{2}$$

Therefore, a qubit represents the conjugated states with the complex numbers α and β, defined as (3):

$$|\psi\rangle = \begin{pmatrix}\alpha\\\beta\end{pmatrix} = \alpha|0\rangle + \beta|1\rangle \tag{3}$$

If α and β are not null, we could describe these factors as the 0 or 1 probability of the state representing a superposition as (4):

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle = \alpha\frac{|+\rangle + |-\rangle}{\sqrt{2}} + \beta\frac{|+\rangle - |-\rangle}{\sqrt{2}} \tag{4}$$

This means that unlike the classic bits, qubits can have both states at the same time (0 and 1). On the other hand, quantum entanglement only occurs between two or more qubits generating a unique state of the system. This intrinsic characteristic of the particles, without similarities in classical theories, is known as the ERP paradox due to its prediction in 1935 by Einstein, Podolsky and Rosen [24]. For our interest in this work, we will only focus on the characteristic of the combination of the quantum states of one or more qubits. The maximum entanglement between two qubits is called the Bell state [25] and its mathematical notation is (5):

$$|\psi\rangle = \frac{|01\rangle - \langle 10|}{\sqrt{2}} \tag{5}$$

where, the possible states of $|\psi\rangle$ are $\{\alpha\,|\,00\rangle, \beta\,|\,01\rangle, \gamma\,|\,10\rangle, \delta\,|\,11\rangle\}$, and α, β, γ, δ are the probabilities each state. This quantum capacity probably increases to infinity depending on the number of entangled qubits. One of the singularities of this formulation is the capacity for the continuity of entanglement of the particles even when they are separated by millions of kilometres. Furthermore, with the feature of parallelism there is the possibility of simultaneously representing the values 0 and 1. Quantum algorithms that operate on superposition states, simultaneously perform operations on all combinations of the inputs. This is where the potential of quantum computers resides.

Another fundamental question in the characteristics of quantum computers is the transformation of the states of the qubits. Currently the technologies used for handling qubits are based on superconducting circuits, ion traps or photonic circuits. These complex physical structures have as their fundamental objective to deliberately modify the states of the qubits. These changes generate the algorithms programmed to yield the desired results of the quantum processors.

The evolution of a closed quantum system is described by a unit transformation [26]. This is that state $|\psi\rangle$ from the system to time t_1 is linked to the state $|\psi'\rangle$ at the time t_1 by a unitary operator U that depends only on t_1 and t_2 such that $|\psi'\rangle = U|\psi\rangle$. A unitary operator U is neither more nor less than a matrix, therefore, applying U on a state is to operate the system by the matrix U. It follows that the state $|\psi'\rangle$ will be determined by the application of a unitary operator (a matrix). As can be seen in Fig. 3, U_f is the unitary operator applied to the state $|\psi\rangle$ and throws us as a result $|\psi'\rangle$:

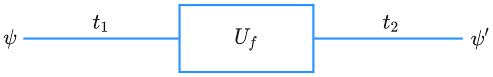


Fig. 3. Unitary operator. Graphical representation of the transformation of states $|\psi\rangle$ on timeline t, when applying a unitary operator U_f.

Therefore, like classical logic gates, state operators modify the states of the qubits, although with some differentiating characteristics over the classical ones. Basically, a quantum gate is a unitary matrix, which, when applied to the qubits, performs a state transformation. The combination of the quantum gates together with the control artifacts generates the unit operators that make up the quantum circuits. Next, we detail the generator of entanglements.

The Hadamard gate can only be applied to a qubit, and its main function on the application of a qubit is for state 0, $H|0\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$, and for state 1, $H|1\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$. It is also defined in matrix form as (6):

$$H = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\1 & -1\end{pmatrix} \tag{6}$$

We can observe in (8) the internal states of a qubit by applying the Hadamard gate to a state $|0\rangle$ and $|1\rangle$. These states (7) are not maintained in the measurement of the qubit, but collapse to 0 or 1.

$$\frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\1 & -1\end{pmatrix}\cdot\begin{pmatrix}0\\1\end{pmatrix} = \begin{pmatrix}1\\-1\end{pmatrix}\,;\ \frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\1 & -1\end{pmatrix}\cdot\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}1\\1\end{pmatrix} \tag{7}$$

This represents the quantum Fourier transformation [27] rotating π about the z axis followed by a rotation of $\frac{\pi}{2}$ about the y axis. In addition to having the characteristic of generating a Bell state, that is, interlacing and deinterlacing qubits in the way described in (8).

$$\begin{array}{lll} |0\rangle & \text{—}\boxed{H}\text{—} & \frac{1}{\sqrt{2}}\ (|0\rangle + |1\rangle) \\ |0\rangle & \text{—}\boxed{H}\text{—} & \frac{1}{\sqrt{2}}\ (|0\rangle + |1\rangle) \end{array} = \frac{1}{2}(|00\rangle + |01\rangle + |10\rangle + |11\rangle) \tag{8}$$

where gate H_4 becomes 4 × 4 matrix represented in (9):

$$\mathrm{H} = \begin{pmatrix}1 & 1 & 1 & 1\\1 & -1 & 1 & -1\\1 & 1 & -1 & -1\\1 & -1 & -1 & 1\end{pmatrix} \tag{9}$$

There are different quantum gates with different objectives or functionalities. For example, the Pauli gate X will exchange the states of a qubit, like the binary NOT gate, that is, if initially the state is $|0\rangle$ it will transform it to a state $|1\rangle$. On the contrary, the Pauli gate Z involves the state $|1\rangle$ exchanging it for -1, leaving the amplitude (probability) of the state $|0\rangle$ untransformed (Fig. 4).

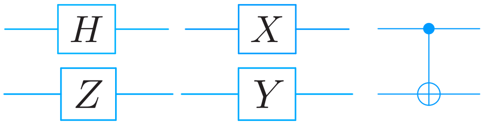


Fig. 4. Graphical representation of quantum gates. H - Hadamard, X - corresponds to a 90° rotation on the x axis, Z corresponds to a 90° rotation on the z axis, Y - corresponds to a 90° rotation on the y axis. In addition to the representation of CNOT.

However, one of the negative characteristics of quantum processors is their fragile stability in state conservation. This conservation of states in a processor is known as quantum coherence, and it can be defined as the conservation of the state of a system in superposition with time. This coherence is physically sensitive to interference from the environment, and can be destroyed by vibrations, electromagnetic disturbances, and other circumstantial disturbances such as sounds, earth vibrations or adverse weather effects. That is, the particle collapses in a state as if it were being measured, losing the multistate characteristic. The destruction of quantum coherence is what is known as *decoherence*.

The coherence time in a quantum computer is essential for the correct implementation and obtaining the expected results. That is why scientific research in the field of quantum computer implementation is vital in advancing in this area. Hadamard gates enhance the intrinsic characteristics of qubits entanglement, therefore, not using them makes a quantum computer become a reversible classical processor.

II. Design of Quantum Games

The conceptual paradigms of quantum computing have been formalized and the parameters established in classical game theory. In addition to demonstrating the potentiality in quantum computational efficiency. Therefore, we are now able to unify both concepts to know the theory of quantum games.

The decision-making strategy is related to the environment and the actors involved. If, on the one hand, on the environment model we can differentiate symmetric and asymmetric games, zero sum, with or without Nash equilibrium and Pareto optimal, combinatorial, perfect information, stochastic or differential. On the other hand, according to its actors, we find cooperative, non-cooperative or multiplayer games, among others. However, not all these classic game theory models can be applicable to the design of circuits and quantum algorithms, given their complexity and lack of research. Consequently, a little explored field of research opens here.

A. Requirements in the Design of Quantum Games

According to DiVincenzo, quantum circuits must have the following requirements for their construction:

- a scalable physical system with well characterized qubits,
- the ability to initialize the state of qubits to 0,
- long relevant decohere times,
- a universal set of reversible quantum gates, and
- a specific measurement capability of the qubit.

We will not discuss requirements 3 and 5, since we consider theoretically that the circuit is stable and without decoherence, in addition to generating an appropriate final measurement. But we will add some specific characteristics of quantum circuits, essential for understanding their design:

- A quantum algorithm is iterative, it is not possible to develop loops.
- Quantum states cannot be cloned, there are no FANOUT circuits that can replicate qubits.
- Quantum circuits are inherently parallel, allowing a function $f(x)$ to be evaluated for multiple values of x simultaneously.

On the other hand, we will need a valid justification for the transformation from classical to quantum algorithms. Well, according to Wu et al. [28], it must be considered that a quantum computer without Hadamard gates is essentially a reversible classical computer and therefore, we cannot consider significant the circuits implemented without these unit matrices of game theory classics. Furthermore, if we add to these requirements and characteristics the three principles of quantum games that they develop [29], we have a technical challenge and an unknown computational complexity, since it is currently unknown how BQP compares with BPP.

Even so, there are laudable attempts to debate the classical models of game theory applied to quantum algorithms, such as the Nash equilibrium, that deserve all the attention given that they are generating the theoretical basis for the evolution of this area.

B. Quantum Nash Equilibrium

The historical example to define the Nash equilibrium in a classical system has been the Prisoner's Dilemma [30] where the desertion of each of the players implies the maximum individual benefit. This game also has the characteristic of establishing an example of pure strategies for each player, since chance does not intervene, and probability is not established. The concept of pure strategy was extensively studied by Antoine Cournot in his work on oligopolies [31] and we can consider it as a particular case of mixed strategies.

However, when implementing the Prisoner's Dilemma game in quantum format according to Van Enk Wu et al. [32], individualistic strategies are eliminated, transforming the game from non-cooperative to cooperative given their retrospective contemplation of both players of their strategies. That is, given the reversibility of the unit matrices, it would be possible to go back and optimize the strategy of both players to obtain the maximum payout. Furthermore, entanglement tends to be considered as a mediated communication [33] or a requirement function according to the theory of abstract economics [34], which does not correspond to the original classical game. The EWL quantification protocol [35] has so far been the most accepted for the quantum transformation of the Nash equilibrium. However, Van Enk Wu contradicts the idiosyncrasy of not preserving the non-cooperative game condition and therefore the elimination of the Nash equilibrium. There is an interesting discussion on this topic in [36].

C. Actors and Their Game Decision Strategies

In a non-cooperative game there are players, the setting and the rules of the game that comprise the strategies and the payouts. These components are defined in standard classical game theory [37] and quantum notation as a tuple $\langle N, \Omega, \mathrm{P}\rangle$ where N is the number of players, Ω_j with $1 \leq j \leq N$ are the strategies of each player, and P_j the payoff function $P\colon \Omega_j \rightarrow R^N$ on each of the strategies. Therefore, the interaction between classical and quantum strategies is hypothetically possible.

This interaction can be seen with the payoff matrix in Table 2. Meyer demonstrated with the Penny-Flip game of two players with the zero-sum strategy, that a quantum strategy will always win over a classical one [38].

TABLE II. Matrix of Payments Between Quantum and Classic Strategies

Type of Actors	Classic	Quantum
Classic	(0,0)	(0,1)
Quantum	(1,0)	(1,1)

At first it might seem that a quantum strategy versus a classical one is always the winner, but Anand and Benjamin [39] demonstrated that a particular classical algorithm, such as the one proposed by Meyer, can beat a quantum one in the Penny-Flip game if it is generalized.

Despite everything argued so far, even in the predictive results of any classical game model, human behaviour breaks the mathematical formalism. Well, game theory assumes unrealistic levels of rationality of its players according to Chen and Hogg. If we translate this reality to the probabilistic results of quantum computers, then the initial payout matrix will change for each strategic decision of the players. The similarity in the decision making of a human player with the results of a quantum computer is demonstrated in their work. Where it seems to be verified that quantum entanglement is a rational human cooperative behaviour with the objective of obtaining the maximum benefit.

III. Quantum Games Implementation Techniques

The transformation from classical to quantum algorithms is a real effort given the requirements seen in Section 2. The applied implementation techniques stand out for their originality as the insertion of tertiary qubits – qutrits [40]. Although the debate centres on the insertion of Hadamard doors. The Hadamard unit matrices entangle the states of the qubits generating multiple overlays if they apply to more than one qubit. It is here where the power of quantum computers lies, and its parallelism seen in Section 3. Therefore, the techniques and results to design quantum games open the interpretive debate of the correct implementation.

Considering the limitations of computers and the types of classes of decision strategy problems seen in Section 1, we can affirm, looking at Fig. 2, that all decision strategy problems of type P can be solved by BPP which in turn are integrated into BQP. Therefore, in principle it seems that all decision strategies in game theory can be solved by a quantum computer. Consequently, there will be a computationally quantum resolution that satisfies the best game strategy for each participant or participants. However, the construction and design of these functions in a quantum computational architecture has yet to be solved for all decision strategy problems.

A summary of implemented quantum games can be found in [3] where the game and the quantum contributions that have been made are defined. In Table III, we expand some of these definitions of [3], including some of the technical characteristics used in their implementations. Some interesting techniques applied to other games are listed as well.

TABLE III. Structure and Implementation of Some Quantum Games

Game	Structure	Implementation
Prisoner's dilemma	Hadamard, without Hadamard, Multiplayer	Mathematical notation, Qiskit (IBM), Various unknown computers
Penny flip	Hadamard, without Hadamard	Mathematical notation, Various unknown computers
Five in a Row - Gomoku	Qutrits	Mathematical notation
Sudoku	Without Hadamard	Python
Poker TH	Hadamard, Multiplayer	Qiskit (IBM)
Bingo	Without Hadamard	Qiskit (IBM)
Monty Hall	Hadamard Qutrits	Mathematical notation
Battle of sexes	Hadamard	Mathematical notation
Rock–scissors–paper	Hadamard	Mathematical notation

Prisoner's Dilemma comprises different variants of the game. However, the most classic one is described by Albert W. Tucker who formalized the game on prison rewards [4]. It belongs to the group of zero-sum non-cooperative games, where the Nash equilibrium is determined according to the Pareto optimal strategy. The original game describes the situation of two participants where they ignore the decisions made by both. Two thieves (Alice and Bob) are caught by the police. Since the police do not have enough evidence to convict them, they propose a deal (summarized in Table IV):

- Alice and Bob confess to the crime -> Alice and Bob are sentenced to 6 years in prison.
- Alice or Bob confess the crime -> Whoever does not confess is sentenced to 10 years in prison and the one who has confessed to 1 year.
- Alice and Bob do not confess the crime -> Alice and Bob are sentenced to 1 year in prison.

TABLE IV. Matrix of Payments Classic Game Prisoner's Dilemma

	Confess	No Confess
Confess	(3, 3)	(-5, 5)
No Confess	(5, -5)	(-1, -1)

All the works on this game are implemented in mathematical notation, although it is worth highlighting the comparison made [30] on human strategies and quantum computers, or in [41] the duality map comparison. Others, such as [42] and [43], focus on classical transitions after analysing quantum decoherence and the null wave function. In addition, in the work [44], the multiplayer version is analysed. And in [44] the Hadamard entanglement gates are not used, although they are used in [45] [46] and [47] where the latter also analyses the applied unit matrices.

Penny flip is related to the flipping of a coin and obtaining heads or tails. However, the quantum strategy is added in player Q. It consists of player P placing a coin head up in an opaque box. After that, they will take turns (Q, then P, then Q) shaking the box or not. P wins when the coin is upside down when the box is opened [38]. This is a zero-sum strategy game for two that could be traditionally analysed using the following matrix reward (Table V).

TABLE V. Matrix of Payments Penny Flip

	NN	NF	FN	FF
N	-1	1	1	-1
F	1	-1	-1	1

The most outstanding works on its implementation and debate are [38] and [45]. However, it is worth highlighting the work [46] where an experiment is carried out intertwining four coins.

Five in a Row is original from Japan and known by different names in other countries. It consists of a 15 × 15 or 19 × 19 matrix where the players alternate in placing their chips on the squares. The winner is the player who manages to form a row, column, or diagonal with k of his chips, where k is the number of cells. The generalization or scalability of the game can be described as (m, n, k), where $m \times n$ will be the dimension of the matrix and k the number of continuous lines to get. In these typical games of the five in a row and Weiqi we can highlight the exotic implementation of [26] with qutrits [38].

Sudoku was popularized in Japan in 1986, although it is proven that the original creator was Leonhard Euler (1707-1783), by establishing the guidelines for the calculation of probabilities to represent a series of numbers without repeating incorporated in The Greco-Latin Square, Euler's Square or Orthogonal Latin Square. This game consists of a 9 × 9 matrix where the decimal numbers except 0 are placed in rows and columns. The challenge is not to repeat any decimal in the same row or column or 3 × 3 sub-matrix. The original matrix is initialized with some numbers that offer clues to start filling in the boxes. This is a single player game, and its initialization tracks must be at least 17 to have a single solution. This popular game is implemented in the high-level Python language without quantum entanglement.

Poker-TH is a multiplayer card game in addition to a mediator player. The mediator player exposes his cards, and the other players must decide the best combination between their cards and those of the mediator. The winning card combination is established from the beginning in a hierarchical range. However, the payoff matrix varies depending on the independent strategies. Each player in his turn bets on his combination, and the other players must accept the bet and continue the game or abandon and win nothing. Therefore, the payoff matrix varies according to the rational independent strategies. It is a game of sum greater than 0, however, in the individual strategies the Nash equilibrium is established, although not in the global game. This game has several variations and the most relevant are based on the number of cards to be dealt between the players and the limitations in the payout matrix. In [48], this game is implemented in IBM quantum computers, with the didactic goal of teaching quantum computing. In addition, it demonstrates decoherence and error mitigation techniques.

Bingo is a very popular game of chance. Players have numbered boxes, and the mediator randomly draws a number from the initially established range. The winning player must own all the numbers in his box from the previously drawn numbers. In this game, the combinatorics, and the number of boxes that each player has intervenes directly. The calculation through the hypergeometric distribution determines the probability (10) of the player to win:

$$P = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \tag{10}$$

where k is the number of the box, x the value of the variable or number of hits in each extraction, N the size of the sample and n the number of each extraction. Therefore, in a range of 90 numbers, the probability of winning on draw 65 is 0.45% for each box. This game is implemented without Hadamard gates of entanglement in the work of [49] on IBM quantum computers.

Monty Hall is taken from the 1975 US television contest *Let's Make a Deal* and has become a real mathematical problem of probability. The name was assigned referring to the presenter of this program and we can also find it as the Monty Hall paradox. The game consists of offering the player the choice of opening three doors. Only one of them hides the desired payment such as a car, a house, money, etc ... depending on the version of the game. Once the door has been chosen by the player, the moderator reduces the choice possibilities to 2 by eliminating a door and again offers the player a new choice. The debate on the game's payout probabilities began in 1990 through the journalistic columns written by Marilyn vos Savant [50], solving and demonstrating the theory that Steve Selvin introduced in 1975 [51]. The controversy lies in the change or not of probabilities. That is, at the beginning the payment of the game has a probability of 1/3 over the 2/3 of losing. However, by eliminating one of the gates we would understand that the probability is now ½ pay and ½ lose. Although if we consider the probabilities assigned at the beginning of the game, the best strategy is to select a new door since the probability of payment continues to be 2/3 (Fig. 5).

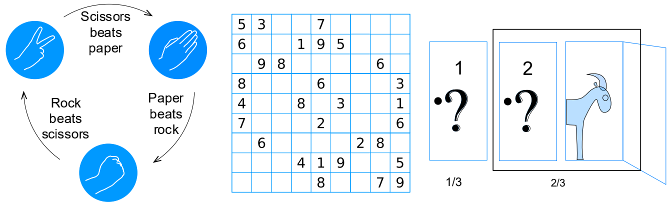


Fig. 5. From left to right: graphical representation of the games Rock-Scissors-Paper, Sudoku, and Monty Hall.

One of the proposals for the implementation of the game in quantum format by [52] is to take the beginning of the game as a three-dimensional Hilbert space where only the moderator reacts in a quantum way leaning on an interlaced Notepad or not, with the aim of saving the information. However, [53] they show that where both participants, contestant and moderator, have access to quantum strategies, the maximum entanglement of the initial states produces the same benefits as the classical game.

Battle of sexes is like the Prisoner's Dilemma, although in this case both participants have full information about payouts and strategies. It consists of selecting the best desired strategy and achieving the greatest benefit individually, therefore, it complies with Nash equilibrium. The initial approach consists of making leisure decisions for a couple (Alice, Bob) with the premise that they both want to be together currently. However, Alice prefers to go to the theatre and Bob wants to go to the movies. The payment matrix according to their preferences is shown in Table VI.

TABLE VI. Matrix of Payments Classic Game Battle of Sexes

		Bob	
		Theatre	**Movie**
Alice	**Theatre**	(2,1)	(0,0)
	Movie	(0,0)	(1,2)

In their article [54], Marinatto and Weber demonstrated that the use of entangled quantum strategies by both players does not improve the classic payoff matrix of the game and therefore generates the same resolution as the classic version of the game. However, a later article by Du *et al.*, [55] discusses [54] approach, proposing a different structure. Thus, Du implemented the game by applying mixed strategies for both players, where each player can freely choose his strategy. That is, they can apply entanglement or not, therefore, the transformation from classical to quantum game seems to demonstrate its efficiency.

Rock-Scissors-Paper was designed in China centuries ago. Today, it is an internationally famous game, and its rules are easy to learn. The game consists of two players simultaneously and in a single movement they determine their non-cooperative strategy. In this case, they determine their weapon which can be rock, paper or scissors. The payoff matrix is represented in Table VII.

TABLE VII. Matrix of Payments Classic Game Rock-Scissors-Paper

	Rock	**Paper**	**Scissors**
Rock	(0, 0)	(-1, 1)	(1, -1)
Paper	(1, -1)	(0, 0)	(-1, 1)
Scissors	(-1, 1)	(1, -1)	(1, 1)

The original game ends in a single action and therefore a Pareto optimal strategy model results. Given that, if one player gets paid, the other gets nothing. However, there is the interpretation of the game with n repetitions. In this case, the game has only one Nash equilibrium and in each round the probability of payout becomes 1/3. Iqbal, in his article [56] studies this game in its repetition format, in an attempt to stabilize the evolutionary sequence (EES) on the Nash equilibrium by applying entanglement to strategies. However, he shows that the odds of winning if both players use quantum strategies are the same as in their classical form.

IV. Conclusion

The implementation of classical quantum algorithms [57] in game theory does not seem to stand solely on the computational techniques used. Quite the contrary, there is a broad debate about changes in strategic models and payment results. Furthermore, in a real decision-making scenario, the intervention of the human being discredits the formulated mathematical formality. Consequently, the predictions of quantum computers seem to be closer to real life scenarios.

It should be considered that the study of the quantum implementation of classical game theory only takes two decades of research. With few published works if we compare them with other disciplines. Therefore, debate and interpretation are still open to great scientific contributions.

Future work will comprehensively address the implementation of binding cooperative and mixed strategy quantum games that include combinatorics.

References

[1] J. Von Neumann and O. Morgenstern, *Theory of games and economic behavior (commemorative edition)*, Princeton university press, 2007.

[2] D. P. DiVincenzo, "The physical implementation of quantum computation," *Fortschritte der Physik Progress of Physics*, vol. 48, no. 9–11, pp. 771–783, 2000.

[3] H. Guo, J. Zhang, and G. J. Koehler, "A survey of quantum games," *Decision Support Systems*, vol. 46, no. 1, pp. 318–332, 2008.

[4] W. Poundstone, *El dilema del prisionero: John Von Neumann, la teoría de juegos y la bomba,* CDU 519.8. , S.A., 1995.

[5] C. E. Shannon, "A chess-playing machine," *Scientific American*, vol. 182, no. 2, pp. 48–51, 1950.

[6] A. M. Turing, "On computable numbers, with an application to the Entscheidungsproblem," *Proceedings of the London Mathematical Society*, vol. 2, no. 1, pp. 230–265, 1937.

[7] A. M. Turing, "Intelligent machinery," NPL. Mathematics Division, 1948.

[8] D. Gallardo López, I. Lesta Pelayo, P. Arques Corrales, *Introducción a la teoría de la computabilidad.* 2003.

[9] J. Hartmanis and R. E. Stearns, "On the computational complexity of algorithms," *Transactions of the American Mathematical Society*, vol. 117, pp. 285–306, 1965.

[10] F. O. F. Llopis E. Pérez, *Programación y Estructura de Datos*, Compobell, SL., 2000.

[11] J. de la Vega, I. Aguilar Juarez, F. Garcia Lamont, and H. Gomez Ayala, "Introduccion al Analisis de Algoritmos," Centro de Estudios e Investigaciones para el Desarrollo Docente, 2019.

[12] R. Solovay and V. Strassen, "A fast Monte-Carlo test for primality," *SIAM Journal on Computing*, vol. 6, no. 1, pp. 84–85, 1977.

[13] P. W. Shor, "Algorithms for quantum computation: discrete logarithms and factoring," in *Proceedings 35th annual symposium on foundations of computer science*, 1994, pp. 124–134.

[14] R. M. Karp, "Reducibility among combinatorial problems," in *Complexity of computer computations*, Springer, 1972, pp. 85–103.

[15] L. J. Stockmeyer, "The polynomial-time hierarchy," *Theoretical Computer Science*, vol. 3, no. 1, pp. 1–22, 1976.

[16] S. Aaronson, "BQP and the polynomial hierarchy," in *Proceedings of the forty-second ACM symposium on Theory of computing*, 2010, pp. 141–150.

[17] Y. Nakata and M. Murao, "Diagonal quantum circuits: their computational power and applications," *The European Physical Journal Plus*, vol. 129, no. 7, pp. 1–14, 2014.

[18] J. Nash, "Non-cooperative games," *Annals of Mathematics*, pp. 286–295, 1951.

[19] C. A. Holt and A. E. Roth, "The Nash equilibrium: A perspective," *Proceedings of the National Academy of Sciences*, vol. 101, no. 12, pp. 3999–4002, 2004.

[20] J. F. Nash and others, "Equilibrium points in n-person games," *Proceedings of the National Academy of Sciences*, vol. 36, no. 1, pp. 48–49, 1950.

[21] V. Pareto, *Cours d'economie politique*, vol. 1. Librairie Droz, 1964.

[22] L. S. Shapley, "Stochastic games," *Proceedings of the National Academy of Sciences*, vol. 39, no. 10, pp. 1095–1100, 1953.

[23] V. Moret, *Principios Fundamentales de la computación cuántica*, 2013.

[24] A. Einstein, B. Podolsky, and N. Rosen, "Can quantum-mechanical description of physical reality be considered complete?," *Physical Review*, vol. 47, no. 10, p. 777, 1935.

[25] J. S. Bell, "On the einstein podolsky rosen paradox," *Physics Physique Fizika*, vol. 1, no. 3, p. 195, 1964.

[26] J. P. Hecht, "*Fundamentos de la computación cuántica orientados a la criptografía teórica*," Editorial Académica Española, 2012.

[27] L. A. Herrera Corredor, "Transformada cuántica de Fourier," *Escuela Colombiana de Ingeniería Julio Garavito*, 2018.

[28] B. Wu, H. Chen, and Z. Luo, "Board Games for Quantum Computers," *arXiv Prepr. arXiv2004.08272*, 2020.

[29] C. F. Lee and N. F. Johnson, "Efficiency and formalism of quantum games," *Physical Review A*, vol. 67, no. 2, p. 22311, 2003.

[30] K.-Y. Chen and T. Hogg, "How well do people play a quantum prisoner's dilemma?," *Quantum Information Processing*, vol. 5, no. 1, pp. 43–67, 2006.

[31] A. A. Cournot, *Recherches sur les principes mathematiques de la theorie des richesses*, L. Hachette, 1838.

[32] S. J. van Enk and R. Pike, "Classical rules in quantum games," *Physical Review A*, vol. 66, no. 2, p. 24306, 2002.

[33] R. J. Aumann, "Subjectivity and correlation in randomized strategies," *Journal of Mathematical Economics*, vol. 1, no. 1, pp. 67–96, 1974.

[34] G. Debreu, "A social equilibrium existence theorem," *Proceedings of the National Academy of Sciences*, vol. 38, no. 10, pp. 886–893, 1952.

[35] J. Eisert, M. Wilkens, and M. Lewenstein, "Quantum games and quantum strategies," *Physical Review Letters*, vol. 83, no. 15, p. 3077, 1999.

[36] F. S. Khan, N. Solmeyer, R. Balu, and T. S. Humble, "Quantum games: a review of the history, current state, and interpretation," *Quantum Information Processing*, vol. 17, no. 11, p. 309, 2018.

[37] M. J. Osborne and A. Rubinstein, *A course in game theory*, MIT press, 1994.

[38] D. A. Meyer, "Quantum strategies," *Physical Review Letters*, vol. 82, no. 5, p. 1052, 1999.

[39] N. Anand and C. Benjamin, "Do quantum strategies always win?," *Quantum Information Processing*, vol. 14, no. 11, pp. 4027–4038, 2015.

[40] Y.-M. Di and H.-R. Wei, "Elementary gates for ternary quantum logic circuit," *arXiv Prepr. arXiv1105.5485*, 2011.

[41] T. Ichikawa and I. Tsutsui, "Duality, phase structures, and dilemmas in symmetric quantum games," *Annals of Physics (N. Y).*, vol. 322, no. 3, pp. 531–551, 2007.

[42] A. Iqbal and A. H. Toor, "Quantum repeated games," *Physics Letters A*, vol. 300, no. 6, pp. 541–546, 2002.

[43] A. Iqbal, "Quantum games with a multi-slit electron diffraction setup," *arXiv Prepr. quant-ph/0207078*, 2002.

[44] J. Du, X. Xu, H. Li, X. Zhou, and R. Han, "Playing prisoner's dilemma with quantum rules," *Fluctuation and Noise Letters*, vol. 2, no. 04, pp. R189--R203, 2002.

[45] E. W. Piotrowski and J. Sładkowski, "An invitation to quantum game theory," *International Journal of Theoretical Physics*, vol. 42, no. 5, pp. 1089–1099, 2003.

[46] A. Iqbal, "Playing games with EPR-type experiments," *arXiv Prepr. quant-ph/0507152*, 2005.

[47] A. Pal, S. Chandra, V. Mongia, B. K. Behera, and P. K. Panigrahi, "Solving Sudoku Game Using Quantum Computation," 2018, doi: 10.13140/RG. 2.2.

[48] F. G. Fuchs, V. Falch, and C. Johnsen, "Quantum Poker—a game for quantum computers suitable for benchmarking error mitigation techniques on NISQ devices," *The European Physical Journal Plus*, vol. 135, no. 4, p. 353, 2020.

[49] V. Singh, B. K. Behera, and P. K. Panigrahi, "Design of quantum circuits to play bingo game in a quantum computer," 2019, doi: https://doi.org/10.13140/RG.

[50] J. S. Rosenthal, "Monty hall, monty fall, monty crawl," *Math Horizons*, vol. 16, no. 1, pp. 5–7, 2008.

[51] S. Selvin, "Monty Hall Problem," *American Statistician*, vol. 29, no. 3, p. 134, 1975.

[52] G. M. D'ariano, R. D. Gill, M. Keyl, B. Kuemmerer, H. Maassen, and R. F. Werner, "The quantum monty hall problem," *arXiv Prepr. quant-ph/0202120*, 2002.

[53] A. P. Flitney and D. Abbott, "Quantum version of the Monty Hall problem," *Physical Review A*, vol. 65, no. 6, p. 62318, 2002.

[54] L. Marinatto and T. Weber, "A quantum approach to static games of complete information," *Physics Letters A*, vol. 272, no. 5–6, pp. 291–303, 2000.

[55] J. Du, H. Li, X. Xu, M. Shi, X. Zhou, and R. Han, "Remark on quantum battle of the sexes game," *arXiv Prepr. quant-ph/0103004*, 2001.

[56] A. Iqbal, "Studies in the theory of quantum games," *arXiv Prepr. quant-ph/0503176*, 2005.

[57] R. Pérez-Anton, A. Corbi, J. I. López-Sánchez, and D. Burgos, "Reliability of IBM's Public Quantum Computers", International Journal of Interactive Multimedia and Artificial Intelligence, In Press, https://doi.org/10.9781/ijimai.2023.04.005.

Raquel Pérez-Antón

Raquel Pérez-Antón is a PhD Candidate in computer science at the Universidad Internacional de La Rioja (UNIR), Graduated in Computer Engineering from the Universidad Internacional de La Rioja (UNIR), Technical Engineer in Management Computer Science from the Universidad de Alicante (UA), 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.

José Ignacio López Sánchez

Dr. José Ignacio López 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 (Fundación 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, https://www.unir.net/facultades/esit/), part of the Universidad Internacional de La Rioja (UNIR, https://www.unir.net/), 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. He also attends and exhibits regularly at international conferences.

Dr. Alberto Corbi

Dr. Alberto Corbi obtained his PhD in Physics at the Universidad de Valencia (UV) and the Institute for Corpuscular Physics (belonging to the 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 an assistant professor at the Engineering School, which are both part of the Universidad Internacional de La Rioja (UNIR). With a background in ocean-atmosphere interaction (M.Sc. obtained at the Universidad Católica de Valencia – UCV), he is currently involved in a variety of research fields: eLearning standards, systems interoperability, server-sided programming languages and solutions medical physics, radiological protection/survey, education-enhanced science education, monitoring of physical activities through inertial sensors (with a special focus on martial arts), social implications of technology (with emphasis on social networks), eHealth advancement (with an accent on Alzheimer's disease characterization and clinical information standards) and environmentalism. 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.

Comprehensive Evaluation of Matrix Factorization Models for Collaborative Filtering Recommender Systems

Jesús Bobadilla, Jorge Dueñas-Lerín, Fernando Ortega, Abraham Gutierrez *

Dpt. Sistemas Informáticos and KNODIS Research Group, Universidad Politécnica de Madrid (Spain)

* Corresponding author: jesus.bobadilla@upm.es (J. Bobadilla), jorgedl@alumnos.upm.es (J. Dueñas-Lerìn), fernando.ortega@upm.es (F. Ortega), abraham.gutierrez@upm.es (A. Gutierrez)..

Received 2 July 2022 | Accepted 4 March 2023 | Early Access 28 April 2023

Abstract

Matrix factorization models are the core of current commercial collaborative filtering Recommender Systems. This paper tested six representative matrix factorization models, using four collaborative filtering datasets. Experiments have tested a variety of accuracy and beyond accuracy quality measures, including prediction, recommendation of ordered and unordered lists, novelty, and diversity. Results show each convenient matrix factorization model attending to their simplicity, the required prediction quality, the necessary recommendation quality, the desired recommendation novelty and diversity, the need to explain recommendations, the adequacy of assigning semantic interpretations to hidden factors, the advisability of recommending to groups of users, and the need to obtain reliability values. To ensure the reproducibility of the experiments, an open framework has been used, and the implementation code is provided.

Keywords

Collaborative Filtering, Matrix Factorization, Recommender Systems.

DOI: 10.9781/ijimai.2023.04.008

I. Introduction

Recommender System (RS) [1] is the field of artificial intelligence specialized in user personalization. Mainly, RSs provide accurate item recommendations to users: movies, trips, books, music, etc. Recommendations are made following some filtering approach. The most accurate filtering approach is the Collaborative Filtering (CF) [2], where recommending to an active user involves a first stage to make predictions about all his or her not consumed or voted items. Then, the top predicted items are recommended to the active user. The CF approach assumes the existence of a dataset that contains explicitly voted items or implicitly consumed items from a large number of users. Remarkable commercial RSs are Amazon, Spotify, Netflix, or TripAdvisor.

Regardless of the machine learning model used to implement CF, the key concept is to extract user and item patterns and then to recommend to the active user those items that he or she has not voted or consumed, and that similar users have highly valued. It fits with the K Nearest Neighbors (KNN) memory-based algorithm [3], and it is the reason why the initial RS research was based on KNN. There are also some other filtering approaches such as demographic, social, content-based, context-aware, and their ensembles. Demographic filtering [4] makes use of user information such as gender, age, or zip code, and item information such as movie genre, country to travel, etc. Social filtering [5], [6] has a growing importance in current RS, due to the social networks boom. The existence of trust relations and graphs [7] can improve the quality of the CF recommendations. In this decentralized and dynamic environment, trust between users provides additional information to the centralized set of ratings. Trust relationships can be local, collective, or global [8]; local information is based on shared users' opinions, collective information uses friends' opinions, whereas global information relates to users' reputation [9]. Content-based filtering [10] recommends items with the same type (content) to consumed items (e.g. to recommend Java books to a programmer that bought some other Java book). Context-aware filtering [11] uses GPS information, biometric sensor data, etc. Finally, ensemble architectures [12] get high accuracy by merging several types of filtering.

Memory-based algorithms have two main drawbacks: their accuracy is not high, and each recommendation process requires to recompute the whole dataset. Model-based approaches solve both problems: their accuracy is higher than that of memory-based methods, and they first create a model from the dataset. From the created model we can make many different recommendations, and it can be efficiently updated when the dataset changes. Matrix Factorization (MF) [13] is the most popular approach to implement current RSs: it provides accurate predictions, it is conceptually simple, it has a straightforward

Please cite this article as: J. Bobadilla, J. Dueñas-Lerín, F. Ortega, A. Gutierrez, "Comprehensive Evaluation of Matrix Factorization Models for Collaborative Filtering Recommender Systems", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 6, pp. 15-23, 2024, http://dx.doi.org/10.9781/ijimai.2023.04.008

implementation, the model learns fast, and also updates efficiently. The MF model makes a compression of information, coding very sparse and large vectors of discrete values (ratings) to low dimensional embeddings of real numbers, called hidden factors. The hidden factors, both from the user vector and from the item vector, are combined by means of a dot product to return predictions. This is an iterative process in which the distance between training predictions and their target ratings is minimized.

The Probabilistic Matrix Factorization (PMF) model based on MF [13] scales linearly with the size of the data set. It also returns accurate results when applied to sparse, large, and imbalanced CF datasets. PMF has also been extended to include an adaptive prior on the model parameters, and it can generalize adequately, providing accurate recommendation to cold-start users. CF RSs are usually biased. A typical CF bias source comes from the fact that some users tend to highly rate items (mainly 4 and 5 stars), whereas some other users tend to be more restrictive in their ratings (mainly 3 and 4 stars). This fact leads to the extension of the MF model to handle biased data. An user-based rating centrality and an item-based rating centrality [14] have been used to improve the accuracy of the regular PMF. These centrality measures are obtained by processing the degree of deviation of each rating in the overall rating distribution of the user and the item. non-Negative Matrix Factorization (NMF) [15] can extract significant features from sparse and non-negative CF datasets (please note that CF ratings are usually a non-negative number of stars, listened songs, watched movies, etc.). When nonnegativity is imposed, prediction errors are reduced and the semantic interpretability of hidden factors is easier. The Bernoulli Matrix Factorization (BeMF) [16] has been designed to provide both prediction and reliability values; this model uses the Bernoulli distribution to implement a set of binary classification approaches. The results of the binary classification are combined by means of an aggregation process. The Bayesian non-Negative Matrix Factorization (BNMF) [17] was designed to provide useful information about user groups, in addition to the PMF prediction results. The authors factorize the rating matrix into two nonnegative matrices whose components lie within the range [0, 1]. The resulting hidden factors provide an understandable probabilistic meaning. Finally, The User Ratings Profile Model (URP) is a generative latent variable model [18]; it produces complete rating user profiles. In the URP model, first attitudes for each item are generated, then a user attitude for the item is selected from the set of existing attitudes. URP borrows several concepts from LDA [19] and the multinomial aspect model [20].

The set of MF models mentioned above: PMF, Biased Matrix Factorization (BiasedMF), NMF, BeMF, BNMF, and URP, can be considered representative in the CF area. These models will be used in this paper to compare their behavior when applied to representative datasets. Specifically, the following quality measures will be tested: Mean Absolute Error (MAE), novelty, diversity, precision, recall, and Normalized Discounted Cumulative Gain (NDCG). Prediction accuracy will be tested using MAE [21], whereas NDCG, Precision and Recall [22] will be used to test recommendation accuracy. Modern CF models should be tested not only regarding accuracy, but also beyond accuracy properties [23]: novelty [24], [25] and diversity [26]. Novelty can be defined as the quality of a system to avoid redundancy; diversity is a quality that helps to cope with ambiguity or under-specification. The models have been tested using four CF datasets: MovieLens (100K and 1M versions) [27], Filmtrust [28] and MyAnimeList [29]. These are representative open datasets and are popular in RS research.

Overall, this paper provides a complete evaluation of MF methods, where the PMF, BiasedMF, NMF, BeMF, BNMF, and URP models have been tested using representative CF quality measures, both for prediction and recommendation, and also beyond accuracy ones. As far as we know this is the experimental most complete work evaluating current MF models in the CF area.

The rest of the paper is structured as follows: Section II introduces the tested models, the experiment design, the selected quality measures, and the chosen datasets. Section III shows the obtained results and provides their explanations in Section IV. Section V highlights the main conclusions of the paper and the suggested future works. Finally, a references section lists current research in the area.

II. Methods and Experiments

This section abstracts the fundamentals of each baseline model (PMF, BiasedMF, NMF, BeMF, BNMF, URP), introduces the tested quality measures (MAE, precision, recall, NDCG, novelty, diversity), and shows the main parameters of the tested datasets (`Movielens`, `FilmTrust`, `MyAnimeList`). Experiments are performed by combining the previous entities.

The vanilla MF [13], [30] is used to generate rating predictions from a matrix of ratings R. This matrix contains the set of casted ratings (explicit or implicit) from a set of users U to a set of items I. Since regular users only vote or consume a very limited subset of the available items, matrix R is very sparse. The MF key concept is to compress the very sparse item and user vectors of ratings to small size and dense item and user vectors of real numbers; these small size dense vectors can be considered as embeddings, and they usually are called ‘hidden factors’, since each embedding factor codes some complex non-lineal (‘hidden’) relation of user or item features. The parameter K is usually chosen to set the embedding (hidden factors) size. MF makes use of two matrices: $P(|U|*K)$ to contain the K hidden factors of each user, and $Q(|I|*K)$ to contain the K hidden factors of each item. To predict how much a user u likes an item i, we compare each hidden factor of u with each corresponding hidden factor of i. Then, the dot product $u \cdot i$ can be used as suitable CF prediction measure. MF predicts ratings by minimizing errors between the original R matrix and the predicted $\hat{R}$ matrix:

$$R \approx P \times Q^T = \hat{R} \tag{1}$$

$$\hat{r}_{ui} = p_u \cdot q_i^T = \sum_{k=1}^{K} p_{uk} \cdot q_{ki} \tag{2}$$

Using gradient descent, we minimize learning errors (differences between real ratings r and predicted ratings $\hat{r}$).

$$e_{ui}^2 = (r_{ui} - \hat{r}_{ui})^2 = \left(r_{ui} - \sum_{k=1}^{K} p_{uk} \cdot q_{ki} \right)^2 \tag{3}$$

To minimize the error, we differentiate equation (3) with respect to p_{uk} and q_{ki}:

$$\frac{\partial}{\partial p_{uk}} e_{ui}^2 = -2(r_{ui} - \hat{r}_{ui}) q_{ki} = -2 e_{ui} q_{ki} \tag{4}$$

$$\frac{\partial}{\partial q_{ki}} e_{ui}^2 = -2(r_{ui} - \hat{r}_{ui}) p_{uk} = -2 e_{ui} p_{uk} \tag{5}$$

Introducing the learning rate α, we can iteratively update the required hidden factors p_{uk} and q_{ki}:

$$p'_{uk} = p_{uk} + \alpha \frac{\partial}{\partial p_{uk}} e_{ui}^2 = p_{uk} + 2\alpha e_{ui} q_{ki} \tag{6}$$

$$q'_{ki} = q_{ki} + \alpha \frac{\partial}{\partial q_{ki}} e_{ui}^2 = q_{ki} + 2\alpha e_{ui} p_{uk} \tag{7}$$

CF datasets have biases, since different users vote or consume items in different ways. In particular, there are users who are more demanding than others when rating products or services. Analogously, there are items more valued than others on average. Biased MF [14] is designed to consider data biases; The following equations extend the previous ones, introducing the bias concept and making the necessary regularization to maintain hidden factor values in their suitable range:

$$\hat{r}_{ui} = \mu + b_u + b_i + p_u \cdot q_i^T \tag{8}$$

where μ, b_u, b_i are the average bias, the user bias and the item bias.

We minimize the regularized squared error:

$$\sum_{r_{ui} \in R_{train}} (r_{ui} - \hat{r}_{ui})^2 + \lambda\left(b_i^2 + b_u^2 + \|q_i\|^2 + \|p_u\|^2\right) \tag{9}$$

where λ is the regularization term.

Obtaining the following updating rules:

$$b_u' \quad = b_u + \alpha(e_{ui} - \lambda b_u) \tag{10}$$

$$b_i' \quad = b_i + \alpha(e_{ui} - \lambda b_i) \tag{11}$$

$$p_u' \quad = p_u + \alpha(e_{ui} \cdot q_i - \lambda p_u) \tag{12}$$

$$q_i' \quad = q_i + \alpha(e_{ui} \cdot p_u - \lambda q_i) \tag{13}$$

NMF [15] can be considered as a regular MF subject to the following constraints:

$$R \geqslant 0, P \geqslant 0, Q \geqslant 0 \tag{14}$$

In the NMF case, predictions are made by linearly combining positive coefficients (hidden factors). NMF hidden factors are easier to semantically interpret than regular MF ones: sometimes it is not straightforward to assign semantic meanings to negative coefficient values. In the CF context, another benefit of using NMF decomposition is the emergence of a natural clustering of users and items. Intuitively, users and items can be clustered according to the dominant factor (i.e. the factor having the highest value). In the same way, the original features (gender, age, item type, item year, etc.) can be grouped according to the factor (from the k hidden factors) on which they have the greatest influence. This is possible due to the condition of positivity of the coefficients.

BeMF [16] is an aggregation-based architecture that combines a set of Bernoulli factorization results to provide pairs <prediction, reliability>. BeMF uses as many Bernoulli factorization processes as possible scores in the dataset. Reliability values can be used to detect shilling attacks, to explain the recommendations, and to improve prediction and recommendation accuracy [31]. BeMF is a classification model based on the Bernoulli distribution. It adequately adapts to the expected binary results of each of the possible scores in the dataset. Using BeMF, the prediction for user u to item i is a vector of probabilities $(p_{ui}^1, \ldots, p_{ui}^D)$, where p_{ui}^s is the probability that i is assigned the s-th score from user u. The BeMF model can be abstracted as follows:

Let $S = \{s_1, \ldots, s_D\}$ be the set of D possible scores in the dataset (e.g. 1 to 5 stars: $D = 5$). From R we generate D distinct matrices $(R^{s1}, \ldots, R^{sD})$; each $R^s = (R_{ui}^s)$ matrix is a sparse matrix such that $R_{ui}^s = 1$. BeMF will attempt to fit the matrices $R^{s1}, \ldots, R^{sD}$ by performing D parallel MFs

The BeMF assumes that, given the user P matrix and the item Q matrix containing $k > 0$ hidden factors, the rate R_{ui} is a Bernoulli distribution with the success probability $\psi(P_u \cdot Q_i)$. The mass function of this random variable is:

$$p(R_{ui} \mid P_u, Q_i) = \begin{cases} \psi(P_u Q_i) & \text{if } R_{ui} = 1 \\ 1 - \psi(P_u Q_i) & \text{if } R_{ui} = 0 \end{cases} \tag{15}$$

The associated likelihood is:

$$\ell(R \mid UV) = \textstyle\sum_{R_{ui}=1} \log\left(\psi(P_u Q_i)\right) + \sum_{R_{ui}=0} \log\left(1 - \psi(P_u Q_i)\right) \tag{16}$$

The BeMF updating equations are:

$$P_u' = P_u + \gamma\left(\sum_{\{i|R_{ui}=1\}} (1 - \text{logit}\ (P_u Q_i))Q_i + \sum_{\{i|R_{ui}=1\}} \text{logit}\ (P_u Q_i)Q_i - \eta P_u\right) \tag{17}$$

$$Q_i' = Q_i + \gamma\left(\sum_{\{i|R_{ui}=1\}} (1 - \text{logit}\ (P_u Q_i))P_u + \sum_{\{i|R_{ui}=1\}} \text{logit}\ (P_u Q_i)P_u - \eta Q_i\right) \tag{18}$$

And the aggregation to obtain the final output Φ:

$$\Phi(u, i) = \frac{1}{\sum_{\alpha=1}^{S} \psi\left(P_u^{S_\alpha} Q_i^{S_\alpha}\right)} \left(\psi\left(P_u^{S_1} Q_i^{S_1}\right), \ldots, \psi\left(P_u^{S_D} Q_i^{S_D}\right)\right) \tag{19}$$

where $\Phi(u, i) = (p_{ui}^1, \ldots, p_{ui}^D)$, $0 \leqslant p_{ui}^\alpha \leqslant 1$, $\sum_\alpha p_{ui}^\alpha = 1$. Let $\alpha_0 = \text{argmax}_\alpha\, p_u^\alpha$; the prediction is: $\hat{R}_{ui} = s_{\alpha_0}$, and the reliability is $p_{ui}^{\alpha_0}$.

BNMF [17] provides a Bayesian-based NMF model that not only allows accurate prediction of user ratings, but also to find groups of users with the same tastes, as well as to explain recommendations. The BNMF model approximates the real posterior distribution $\rho(\emptyset_u, k_{ik}, z_{ui} \mid \rho_{ui})$ by the distribution:

$$q(\emptyset_u, k_{ik}, z_{ui}) = \Pi_{u=1}^{N} q_\emptyset(\emptyset_u) \Pi_{i=1}^{M} \Pi_{k=1}^{K} q_{k_{ik}}(k_{ik}) \\ \Pi_{r_{ui} \neq *} q_{z_{ui}}(z_{ui}) \tag{20}$$

where:

- $z_{ui} \sim \text{Cat}\ (\emptyset_u)$ is a random variable from a categorical distribution.
- $\rho_{ui} \sim \text{Bin}\left(R, k_{i,z_{ui}}\right)$ is a random variable from a Binomial distribution (which takes values from 0 to $D - 1$)
- $p_{ui} = \sum_{k=1 \ldots K} a_{uk} \cdot b_{ik}$ (a and b are hidden matrices).
- $q_{ui} = \begin{cases} 1 & \text{if } 0 \leqslant p_{ui} < 0.2 \\ 2 & \text{if } 0.2 \leqslant p_{ui} < 0.4 \\ \text{etc.} \end{cases}$
- $q_{\emptyset_u}(\emptyset_u) \sim \text{Dir}\ (\gamma_{u1}, \ldots, \gamma_{uk})$ follows a Dirichlet distribution.
- $q_{k_{ik}}(k_{ik}) \sim \text{Beta}\ (\epsilon_{ik}^+, \epsilon_{ik}^-,)$ follows a Beta distribution.
- $q_{z_{ui}}(z_{ui}) \sim \text{Cat}\ (\lambda_{ui1}, \ldots, \lambda_{uik})$ follows a categorical distribution
- λ_{uik} are parameters to be learned: $\lambda_{ui1} + \cdots + \lambda_{uik} = 1$

BNMF iteratively approximates parameters $\{\gamma_{uk}, \epsilon_{ik}^+, \epsilon_{ik}^-, \lambda_{uik}\}$:

$$\gamma_{uk} = \alpha + \sum_{\{i|r_{ui} \neq *\}} \lambda_{uik} \tag{21}$$

$$\epsilon_{ik}^+ = \beta + \sum_{\{i|r_{ui} \neq *\}} \lambda_{uik} \cdot r_{ui}^+ \tag{22}$$

$$\epsilon_{ik}^- = \beta + \sum_{\{i|r_{ui} \neq *\}} \lambda_{uik} \cdot r_{ui}^- \tag{23}$$

$$\lambda_{uik}' = exp(\psi(\gamma_{uk}) + r_{ui}^+ \cdot \psi(\epsilon_{ik}^+) + r_{ui}^- \cdot \psi(\epsilon_{ik}^-) \\ -(D-1) \cdot \psi(\epsilon_{ik}^+ + \epsilon_{ik}^-)) \tag{24}$$

$$r_{ui}^+ = \rho_{ui} = (D-1) \cdot r_{ui}^* \tag{25}$$

$$r_{ui}^- = (D-1) - \rho_{ui} = (D-1) \cdot (1 - r_{ui}^*) \tag{26}$$

$$r_{ui}^* = \frac{\rho_{ui}}{(D-1)} \tag{27}$$

where ψ is the digamma function as the logarithmic derivative of the gamma function.

URP is a generative latent variable model [18]. The model assigns to each user a mixture of user attitudes. Mixing is performed by a

Dirichlet random variable:

$$P(\theta, z | \alpha, \beta, r^u) \approx q(\theta, z \mid \gamma^u, \emptyset_u) = q(\theta \mid \gamma^u) \Pi_{y=1}^{M} q(Z_y = z_y \mid \emptyset_y^u) \tag{28}$$

$$\emptyset_{zy}^u \approx \Pi_{v=1}^{V} \beta_{vyz}^{\delta(r_y^u, v)} \exp\left(\psi(\gamma_z^u) - \psi\left(\sum_{j=1}^{k} \gamma_j^u\right)\right) \tag{29}$$

$$\gamma_z^u = \alpha_z + \sum_{y=1}^{M} \emptyset_{zy}^u \tag{30}$$

$$\beta_{vyz} \approx \sum_{u=1}^{N} \emptyset_{zy}^u \delta(r_y^u, v) \tag{31}$$

$$\psi(\alpha_z) = \psi\left(\sum_{z=1}^{K} \alpha_z\right) + \frac{1}{N} \cdot \left(\sum_{u=1}^{N} \psi(\gamma_z^u) - \psi\left(\sum_{u=1}^{N} \gamma_z^u\right)\right) \tag{32}$$

$$\alpha_z = \psi^{-1}(\psi(\alpha_z)) \tag{33}$$

In this paper, baseline models will be tested using a) prediction measure, b) recommendation measures, and c) beyond accuracy measures. The chosen prediction measure is the MAE, where the absolute differences of the errors are averaged. Absolute precision and relative recall measures are tested to compare the quality of an unordered list of N recommendations. The ordered lists of recommendations will be compared using the NDCG quality measure. From the beyond accuracy metrics, we have selected novelty and diversity. Novelty returns the distance from the items the user 'knows' (has voted or consumed) to his recommended set of items. Diversity tells us about the distance between the set of recommended items. Recommendations with high novelty values are valuable, since they show to the user unknown types of items. Diverse recommendations are valuable because they provide different types of items (and each type of item can be novel, or not, to the user).

The GroupLens research group [27] made available several CF datasets, collected over different intervals of time. MovieLens 100K and MovieLens 1M describe 5-star rating and free-text tagging activity. These data were created from 1996 to 2018. In the Movielens 100K dataset, users were selected at random from those who had rated at least 20 movies, whereas the MovieLens 1M dataset has not this constraint. Only movies with at least one rating or tag are included in the dataset. No demographic information is included. Each user is represented by an 'id', and no other information is provided. The dataset files are written as comma-separated values files with a single header row. Columns that contain commas (,) *areescapedusing* double-quotes ("). These files are encoded as UTF-8. All ratings are contained in the file named 'ratings.csv'. Each line of this file after the header row represents one rating of one movie by one user, and has the following format: 'userId, movieId, rating, timestamp'. The lines within this file are ordered first by 'userId', then, within user, by 'movieId'. Timestamps represent seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970. FilmTrust is a small dataset crawled from the entire FilmTrust website in June, 2011. As the Movielens datasets, it contains ratings voted from users to items; additionally, it provides social information structured as a graph network. Finally, MyAnimeList contains information about anime and 'otaku' consumers (anime, manga, video games and computers). Each user is able to add 'animes' to their completed list and give them a rating; this data set is a compilation of those ratings. The MyAnimeList CF information is contained in the file 'Anime.csv', where their main columns are 'anime_id': myanimelist.net's unique 'id' identifying an anime; 'name': full name of anime; 'genre': comma separated list of genres for this anime; 'type': movie, TV, OVA, etc; 'episodes': how many episodes in this show; 'rating': average rating out of 10 for this anime. These datasets are available in the Kaggle and GitHub repositories, as well as in the KNODIS research group CF4J [32] repository https://github.com/ferortega/cf4j.

Table I contains the values of the main parameters of the selected CF data sets: Movielens 100K, Movielens 1M, FilmTrust and MyAnimeList. We have run the explained MF models on each of the four Table I datasets, testing the chosen quality measures. Please note that the MyAnimeList dataset ratings range from 1 to 10 , whereas MovieLens datasets range from 1 to 5 and FilmTrust ranges from 0 to 5 with 0.5 increments. It is also remarkable the sparsity difference between FilmTrust and the rest of the tested datasets.

TABLE I. Main Parameter Values of the Tested Datasets

Dataset	#users	#items	#ratings	Scores	Sparsity
MovieLens100k	943	1682	99,831	1 to 5	93.71
MovieLens1M	6,040	3,706	911,031	1 to 5	95.94
MyAnimeList	19,179	2,692	548,967	1 to 10	98.94
FilmTrust	1,508	2,071	35,497	0 to 5	87.98

Experiments have been performed using random search and applying four-fold cross-validation. To ensure reproducibility, we used a seed in the random process. Results shown in the paper are the average of the partial results obtained by setting the number *k* of latent factors to {4, 8, 12}, and the number of MF iterations to {20, 50, 75, 100}. Additionally, to run the PMF, BiasedMF, and BeMF models, both the learning rate and the regularization parameters have been set to {0.001, 0.01, 0.1, 1.0}. The BNMF model requires two specific parameters: α and β; the chosen values por these parameters are: α = {0.2, 0.4, 0.6, 0.8}, and β = {5, 15, 25}. The tested number of recommendations N ranges from 1 to 10. We have used 4 stars as recommendation threshold θ for datasets whose ratings range from 1 to 5 , while the testing threshold has been 8 when MyAnimeList was chosen. The experiments have been implemented using the open framework [33] and the code has been made available at https://github.com/KNODIS-Research-Group/choice-of-mf-models.

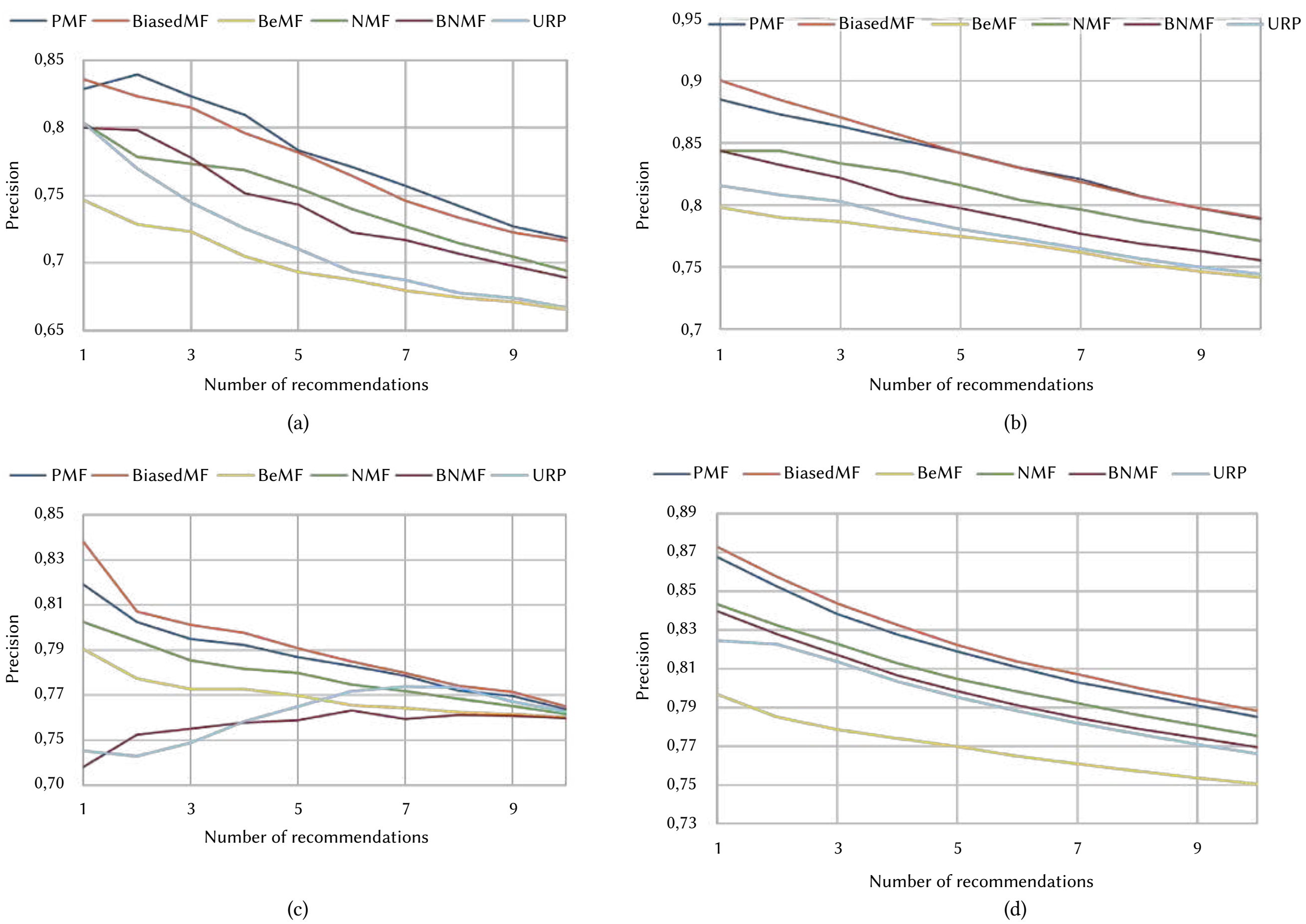
III. Results

The prediction quality obtained by testing each baseline model is shown in table II. The bold numbers correspond to the best results, and, of them, those highlighted gray are the top ones. As can be seen, BiasedMF and BNMF models provide the best CF prediction results. PMF, NMF, BeMF and URP seem to be more sensitive to the type of CF input data.

TABLE II. Prediction Quality Results Using the Mean Absolute Error (MAE). The Lower the Error Value, the Better the Result

	PMF	BiasedMF	NMF	BeMF	BNMF	URP
MovieLens 100K	0.770	**0.754**	0.804	0.805	**0.748**	0.837
MovieLens 1M	0.729	**0.712**	0.744	0.748	**0.693**	0.795
FilmTrust	0.863	**0.652**	0.876	0.712	**0.666**	0.831
MyAnimeList	1.110	**0.926**	1.147	1.034	**0.943**	1.159

Fig. 1 shows the quality of recommendation obtained using the Precision measure. The most remarkable in Fig. 1 is the superiority of the models PMF and BiasedMF. For the remaining models, URP and BeMF provide the worst results, whereas the nonnegative NMF and BNMF return an intermediate quality. It is important to highlight the good performance of the BiasedMF model for both the prediction and the recommendation tasks.



(a)

(b)

(c)

(d)

Fig. 1. Precision recommendation quality results; a) `MovieLens100K`, b) `MovieLens 1M`, c) `FilmTrust`, d) `MyAnimeList`. The higher the values, the better the results.

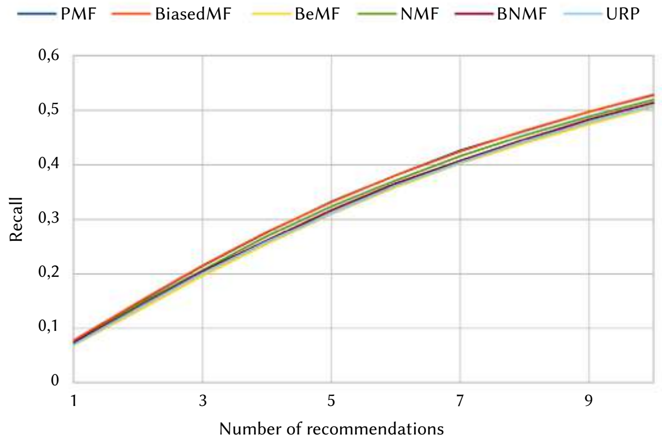


Fig. 2. Recall Recommendation quality results obtained in the `MovieLens` 1M dataset. The results of the other three considered datasets are very similar to this one; to maintain the paper as short as possible, the results of other datasets are not shown.

To test the quality of CF recommendations of unordered recommendations, precision and recall measures are usually processed, and they are provided separately, or joined in the F1 score. We have done these experiments and we have not found appreciable differences in Recall values for the tested models in the selected datasets. In order to maintain the paper as short as possible, Fig. 2 only shows the Recall results obtained by processing the `Movielens 1M` dataset. Results from the rest of datasets are very similar; consequently, the Recall quality measure does not help, in this context, to find out the best MF models in the CF area.

Fig. 2. Recall Recommendation quality results obtained in the MovieLens 1M dataset. The results of the other three considered datasets are very similar to this one; to maintain the paper as short as possible, the results of other datasets are not shown.

In the RSs field, recommendations are usually provided in an ordered list. Users' trust in RSs quickly decays when the first recommendations in the list do not meet their expectations; for that reason, the NDCG quality measure particularly penalizes errors in the first recommendations of the list. Fig. 3 (NDCG results) shows a similar behavior to Fig. 1, where the BiasedMF and PMF models provide the best recommendation quality. So, these two models perform fine both in recommending ordered and unordered lists.

Traditionally, RSs have been evaluated attending to their prediction and recommendation accuracy; nevertheless, there are some other valuable beyond accuracy aims and their corresponding quality measures. The Diversity measure tests the variety of recommendations, penalizing recommendations focused on the same 'area' (Star Wars III, Star Wars I, Star Wars V, Han Solo). Fig. 4 shows the Diversity results obtained by testing the selected models; the most diverse recommendations are usually returned when the BiasedMF model is used, followed by both PMF and NMF. This fact is particularly interesting, since it is not intuitive that the same model (BiasedMF) can, simultaneously, provide accurate and diverse recommendations.

Novelty is an important beyond accuracy objective in RSs. Users appreciate accurate recommendations, but they also want to discover

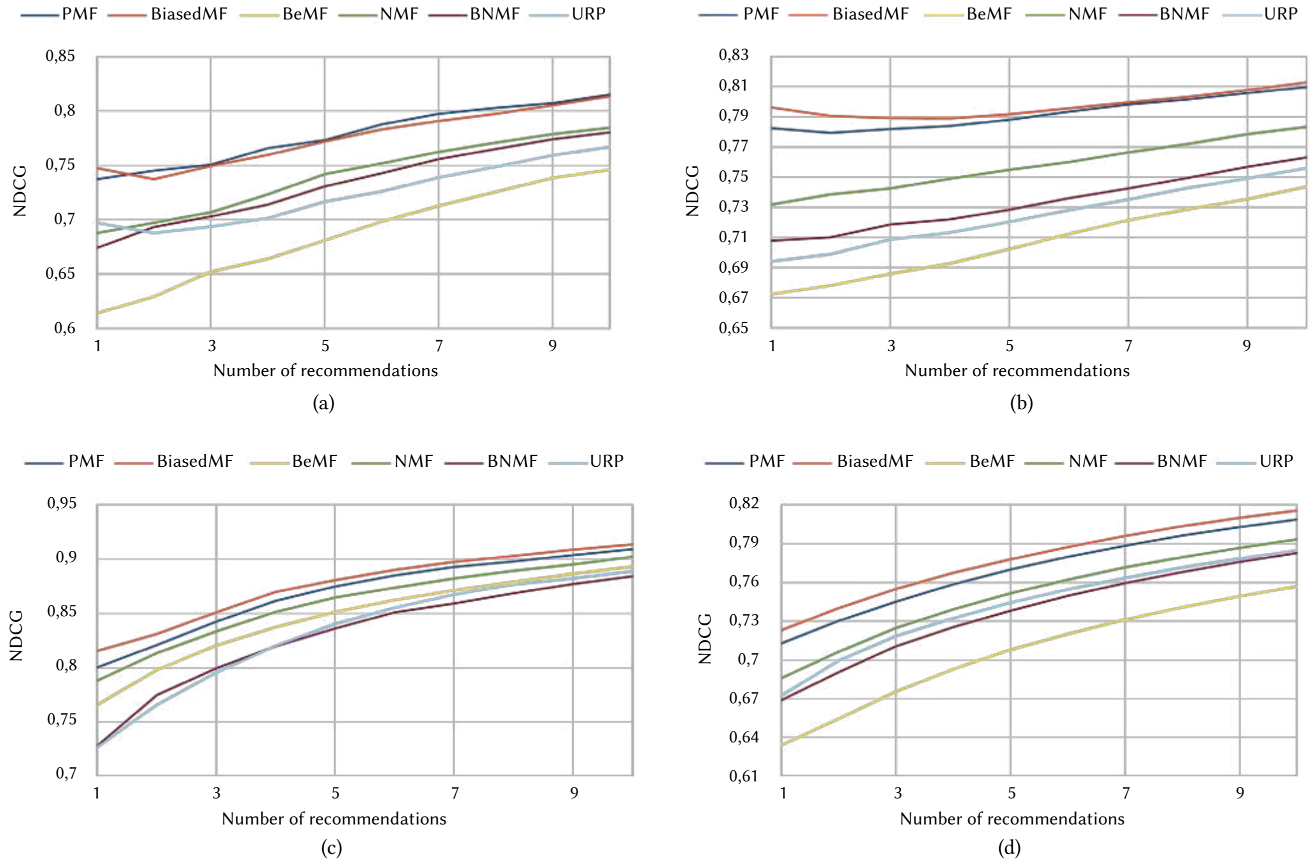


Fig. 3. Normalized Discounted Cumulative Gain recommendation quality results; a) `MovieLens100K`, b) `MovieLens 1M`, c) `FilmTrust`, d) `MyAnimeList`. The higher the values, the better the results.

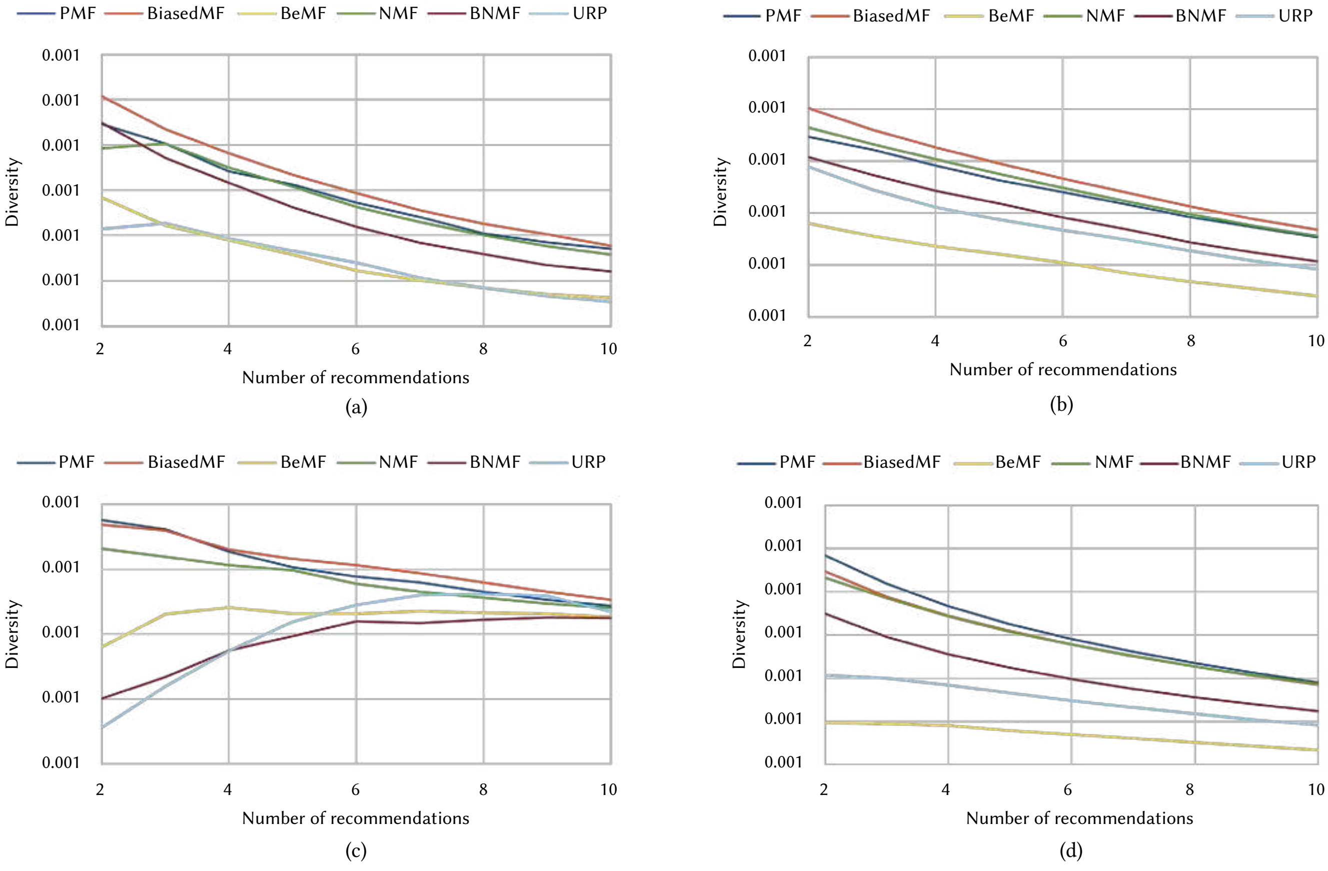
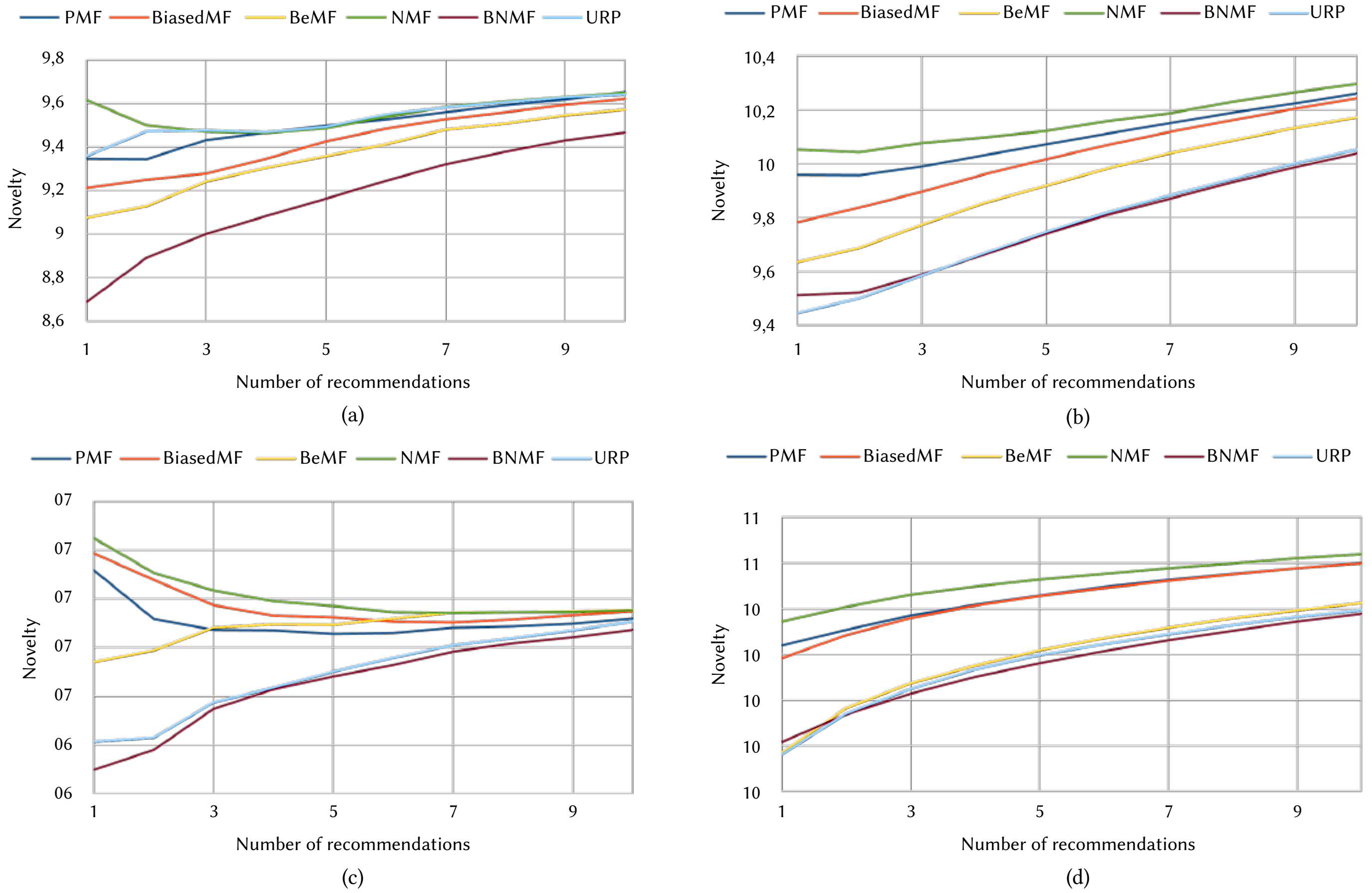


Fig. 4. Diversity beyond accuracy results; a) `MovieLens100K`, b) `MovieLens 1M`, c) `FilmTrust`, d) `MyAnimeList`. The higher the values, the better the results.



(a)

(b)

(c)

(d)

Fig. 5. Novelty beyond accuracy quality results; a) `MovieLens100K`, b) `MovieLens 1M`, c) `FilmTrust`, d) `MyAnimeList`. The higher the values, the better the results.

unexpected (and accurate enough) recommendations. Please note that a set of recommendations can be diverse and not novel, as they can be novel and not diverse. It would be great to receive, simultaneously, accurate, novel, and diverse recommendations, but usually improving some of the objectives leads to worsening others. Fig. 5 shows the results of the novelty quality measure: NMF returns novel recommendations, compared to other models; NMF provides a balance between accuracy and novelty. BiasedMF and PMF also provide novel recommendations compared to BeMF and URP.

IV. Discussion

In this section, we provide a comparative discussion of the most adequate MF models when applied to a set of different CF databases. To judge each MF model, we simultaneously measure a set of conflicting goals: prediction accuracy, recommendation accuracy (unordered and ordered lists) and beyond accuracy aims. We will promote some MF models as ‘winners’, attending to their high performance (overall quality results) when applied to the tested datasets. We also provide a summary table to better identify those MF models that perform particularly fine on any individual quality objective: novelty, diversity, precision, etc., as well as any combination of those quality measures.

Table III summarizes the results of this section. BiasedMF is the most appropriate model when novelty of recommendations is not a particularly relevant issue. PMF can be used instead BiasedMF when simplicity is required (e.g. educational environments). BeMF should only be used when reliability information is required or when reliability values are used to improve accuracy [31]. NMF and BNMF are adequate when semantic interpretation of hidden factors is needed. NMF is the best choice when we want to be recommended with novel items. BNMF provides good accuracy and it is designed to recommend to group of users.

TABLE III. MF Models Comparative

	PMF	BiasedMF	NMF	BeMF	BNMF	URP
MAE	++	+++	+	+	+++	+
Precision	+++	+++	++	+	++	+
NDCG	+++	+++	+	+	+	+
Diversity	++	+++	++	+	+	+
Novelty	++	++	+++	+	+	+
Total	12	14	9	5	8	5

V. Conclusions

This paper makes a comparative of relevant MF models applied to collaborative filtering recommender systems. Prediction, recommendation, and beyond accuracy quality measures have been tested on four representative datasets. The results show the superiority of the BiasedMF model, followed by the PMF one. BiasedMF arises as the most convenient model when novelty is not a particularly important feature. PMF combines simplicity with accuracy; it can be the best choice for educational or not commercial implementations. NMF and BNMF are adequate when we want to do a semantic interpretation of their non-negative hidden factors. NMF is preferable to BNMF when beyond accuracy (novelty and diversity) results are required, whereas it is better to make use of BNMF when prediction accuracy is required or when recommending to group of users, or when explaining recommendations is needed. NMF and BiasedMF are the best choices when beyond accuracy aims are selected, whereas PMF or BiasedMF performs particularly well in recommendation task, both for unordered and ordered options. BeMF can only be selected when reliability values are required or when they are used to improve accuracy. Finally, URP does not seem to be an adequate choice in any of the combinations tested. As future work, it is proposed to add new MF models, quality measures, and datasets to the experiments, as well as the possibility of including neural network models such as DeepMF or Neural Collaborative Filtering (NCF).

Acknowledgments

This work has been co-funded by the *Ministerio de Ciencia e Innovación* of Spain and European Regional Development Fund (FEDER) under grants 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] Z. Batmaz, A. Yurekli, A. Bilge, C. Kaleli, “A review on deep learning for recommender systems: challenges and remedies,” *Artificial Intelligence Review*, vol. 52, no. 1, pp. 1–37, 2019.

[2] J. Bobadilla, S. Alonso, A. Hernando, “Deep learning architecture for collaborative filtering recommender systems,” Applied Sciences, vol. 10, no. 7, p. 2441, 2020.

[3] 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.

[4] J. Bobadilla, R. Lara-Cabrera, Á. González-Prieto, F. Ortega, “Deepfair: Deep learning for improving fairness in recommender systems,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 86–94, 2021, doi: 10.9781/ijimai.2020.11.001.

[5] J. Carbó, J. M. Molina, J. Dávila, “Fuzzy referral based cooperation in social networks of agents,” *AI Communications*, vol. 18, pp. 1–13, 2005. 1.

[6] D. Medel, C. González-González, S. V. Aciar, “Social relations and methods in recommender systems: A systematic review,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, p. 7, 2022, doi: 10.9781/ijimai.2021.12.004.

[7] M. Caro-Martínez, G. Jiménez-Díaz, J. A. Recio- García, “Local model-agnostic explanations for black-box recommender systems using interaction graphs and link prediction techniques,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. InPress, no. InPress, p. 1, 2021, doi: 10.9781/ijimai.2021.12.001.

[8] S. Afef, Z. Brahmi, M. Gammoudi, “Trust-based recommender systems: An overview,” in *27th IBIMA Conference*, 05 2016.

[9] I. Pinyol, J. Sabater-Mir, “Computational trust and reputation models for open multi-agent systems: a review,” *Artificial Intelligence Review*, vol. 40, pp. 1–25, Jun 2013, doi: 10.1007/s10462-011-9277-z.

[10] 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.

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

[12] S. Forouzandeh, K. Berahmand, M. Rostami, “Presentation of a recommender system with ensemble learning and graph embedding: a case on movielens,” *Multimedia Tools and Applications*, vol. 80, no. 5, pp. 7805–7832, 2021.

[13] R. Salakhutdinov, A. Mnih, “Probabilistic matrix factorization,” in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS’07, Red Hook, NY, USA, 2007, p. 1257–1264, Curran Associates Inc.

[14] Z. Wu, H. Tian, X. Zhu, S. Wang, “Optimization matrix factorization recommendation algorithm based on rating centrality,” in *International Conference on Data Mining and Big Data*, 2018, pp. 114–125, Springer.

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

[16] F. Ortega, R. Lara-Cabrera, Á. González-Prieto, J. Bobadilla, “Providing reliability in recommender systems through bernoulli matrix factorization,” *Information Sciences*, vol. 553, pp. 110–128, 2021.

[17] A. Hernando, J. Bobadilla, F. Ortega, “A non negative matrix factorization for collaborative filtering recommender systems based on a bayesian probabilistic model,” *Knowledge-Based Systems*, vol. 97, pp. 188–202, 2016.

[18] B. M. Marlin, “Modeling user rating profiles for collaborative filtering,” *Advances in neural information processing systems*, vol. 16, 2003.

[19] D. M. Blei, A. Y. Ng, M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[20] T. Hofmann, “Learning what people (don’t) want,” in *European Conference on Machine Learning*, 2001, pp. 214– 225, Springer.

[21] A. Gunawardana, G. Shani, “Evaluating recommender systems,” in *Recommender systems handbook*, Springer, 2015, pp. 265–308.

[22] C. C. Aggarwal, “Evaluating recommender systems,” in *Recommender systems*, Springer, 2016, pp. 225–254.

[23] J. Bobadilla, A. Gutiérrez, S. Alonso, Á. 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, doi: 10.9781/ijimai.2021.08.010.

[24] S. Vargas, P. Castells, “Rank and relevance in novelty and diversity metrics for recommender systems,” in *Proceedings of the fifth ACM conference on Recommender systems*, 2011, pp. 109–116.

[25] P. Castells, S. Vargas, J. Wang, “Novelty and diversity metrics for recommender systems: choice, discovery and relevance,” in *Proceedings of the 33rd European Conference on Information Retrieval (ECIR’11)*, 2011.

[26] S. Vargas, P. Castells, D. Vallet, “Intent-oriented diversity in recommender systems,” in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 2011, pp. 1211– 1212.

[27] F. M. Harper, J. A. Konstan, “The movielens datasets: History and context,” *Acm transactions on interactive intelligent systems (tiis)*, vol. 5, no. 4, pp. 1–19, 2015, doi: https://doi.org/10.1145/2827872.

[28] J. Golbeck, J. A. Hendler, “Filmtrust: movie recommendations using trust in web-based social networks,” *CCNC 2006. 2006 3rd IEEE Consumer Communications and Networking Conference, 2006.*, vol. 1, pp. 282–286, 2006, doi: 10.1109/CCNC.2006.1593032.

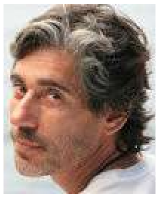
[29] J. Miller, G. Southern, “Recommender system for animated video,” *Issues in Information Systems*, vol. 15, no. 2, pp. 321–7, 2014.

[30] Y. Koren, R. Bell, C. Volinsky, “Matrix factorization techniques for recommender systems,” *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[31] J. Bobadilla, A. Gutiérrez, S. Alonso, Á. 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, doi: 10.9781/ijimai.2021.08.010.

[32] F. Ortega, B. Zhu, J. Bobadilla, A. Hernando, “Cf4j: Collaborative filtering for java,” *Knowledge- Based Systems*, vol. 152, pp. 94–99, 2018, doi: https://doi.org/10.1016/j.knosys.2018.04.008.

[33] F. Ortega, J. Mayor, D. López-Fernández, R. Lara- Cabrera, “Cf4j 2.0: Adapting collaborative filtering for java to new challenges of collaborative filtering based recommender systems,” *Knowledge-Based Systems*, vol. 215, p. 106629, 2021.



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 Applied Intelligent 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 teamworking 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.

Jorge Dueñas-Lerín

Jorge Dueñas-Lerín received the B.S. in computer science from the Universidad Politécnica de Madrid. He received the M.S. degree in highschool, vocational training and languages teacher from the Universidad Nacional de Educación a Distancia. He is currently a Ph.D. student as part of the KNOledge Discovery and Information Systems - KNODIS research group.

Fernando Ortega

Fernando Ortega was born in Madrid, Spain, in 1988. He received the B.S. degree in software engineering, the M.S. degree in artificial intelligence, and the Ph.D. degree in computer sciences from theUniversidad Politécnica de Madrid, in 2010, 2011, and 2015, respectively. He is currently Associate Professor in the Universidad Politécnica de Madrid. He is author of more than 50 research papers in most prestigious international journals. He leads several national projects to include machine learning algorithms into the society. His research interests include machine learning, data analysis, and artificial intelligence. He is the head researcher of the KNOledge Discovery and Information Systems - KNODIS 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.

OBOE: an Explainable Text Classification Framework

Raúl A. del Águila Escobar[1], Mari Carmen Suárez-Figueroa[2], Mariano Fernández-López[3] *

[1] Universidad Politécnica de Madrid (UPM), Boadilla del Monte (Spain)
[2] Ontology Engineering Group (OEG), Universidad Politécnica de Madrid (UPM), Boadilla del Monte (Spain)
[3] Department of Information Technology, Escuela Politécnica Superior, Universidad CEU-San Pablo, Boadilla del Monte (Spain)

* Corresponding author: r.delaguila@alumnos.upm.es (R. A. del Águila Escobar), mcsuarez@fi.upm.es (M. C. Suárez-Figueroa), mfernandez.eps@ceu.es (M. Fernández-López).

Received 18 October 2021 | Accepted 27 September 2022 | Early Access 3 November 2022

LA UNIVERSIDAD EN INTERNET

Abstract

Explainable Artificial Intelligence (XAI) has recently gained visibility as one of the main topics of Artificial Intelligence research due to, among others, the need to provide a meaningful justification of the reasons behind the decision of black-box algorithms. Current approaches are based on model agnostic or ad-hoc solutions and, although there are frameworks that define workflows to generate meaningful explanations, a text classification framework that provides such explanations considering the different ingredients involved in the classification process (data, model, explanations, and users) is still missing. With the intention of covering this research gap, in this paper we present a text classification framework called OBOE (explanatiOns Based On concEpts), in which such ingredients play an active role to open the black-box. OBOE defines different components whose implementation can be customized and, thus, explanations are adapted to specific contexts. We also provide a tailored implementation to show the customization capability of OBOE. Additionally, we performed (a) a validation of the implemented framework to evaluate the performance using different corpora and (b) a user-based evaluation of the explanations provided by OBOE. The latter evaluation shows that the explanations generated in natural language express the reason for the classification results in a way that is comprehensible to non-technical users.

Keywords

Explainable Artificial Intelligence, Explanation, Framework, Text Classification.

DOI: 10.9781/ijimai.2022.11.001

I. Introduction

As a consequence of the wide use of black-box algorithms and the need to provide the justification that supports a classification result, eXplainable Artificial Intelligence (XAI), set up as an initiative, is one of the most relevant research topics in the last years.

Conceptually, a text classification problem is no different from other classification problems, so the same ingredients are involved in solving the problem: data, model, users (final users or model developers) and the context of the classification problem. Therefore, the challenges and questions that text classification tries to answer from an XAI perspective are the same: the need to specify the reasons behind the decision of the model (why question), the context of the explanation (what for question), how the model arrived at a conclusion (how question) or the data and problem of the classification (what question). However, all these ingredients and questions are not being considering together in a system to provide meaningful explanations [1].

The aim of this paper is twofold. Firstly, we present a customizable framework called OBOE (explanatiOns Based On concEpts) for explaining classification of texts. This framework defines a workflow that can be customized and allows all the ingredients to play an active role in the classification process. Furthermore, these ingredients work together to answer the questions that allow the black-box to be opened for final users and model developers by defining the following features: (a) Explanation Generation Workflow (how, why): there is an explicit and defined workflow for generating meaningful explanations for the users; (b) Data as key ingredient (what question): data is not considered just an element used by the Machine Learning (ML) algorithm to classify or by an explainer to provide an explanation. The role of data is to drive the workflow both to classify documents and to obtain meaningful explanations; (c) Agnostic Model (why and how questions): this means that the approach can be used with different ML models in order to be able to choose the one that best suits the problem; (d) External Knowledge Integration (what question): additional resources, such as thesauri or ontologies, are used to enrich explanations; (e) Involvement of Users (what for, why and what): users are actively involved in the process of generating explanations; (f) Explanations in Natural Language (what for question): general purpose explanations are created in natural language; (g) Explanations based on Relevance of Terms (why question): the relevance of terms that appear in the vocabulary defined by the model developer is used to generate the explanations; and (h) Interchangeable Components:

Please cite this article as:
R. A. del Águila Escobar, M. C. Suárez-Figueroa, M. Fernández-López, "OBOE: an Explainable Text Classification Framework", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 6, pp. 24-37, 2024, http://dx.doi.org/10.9781/ijimai.2022.11.001

every component in the approach can be substituted by others performing a similar function.

Secondly, we introduce our first attempt to automatically generate explanations that (a) inform and help users to understand why a particular result is produced by a classification system, and (b) are easy to understand by the so-called final users, who have no technical knowledge. Apart from final or non-technical users, our work considers engineers or scientists who parametrize the model and the workflow (the so-called model developers [2]). Model developers and final users can interact to obtain custom explanations by, for example, choosing the vocabulary to be used in the explanations.

The rest of the paper is structured as follows: in Section II we provide an overview of the literature in the field of XAI. Section III describes the different components of the OBOE framework. In Section IV we present an example of implementation of the framework. Section V presents the validation performed over this implementation, the corpora used, and the results obtained; while in Section VI we explain the user-based evaluation performed over the explanations provided by OBOE as well as the main results of such an evaluation. Section VII exposes, as a recapitulation, the main contributions, and results of our work. Finally, Section VIII provides the conclusions and future lines of work.

II. Related Work

Text classification is still an open research task due to its inner complexity, the real-world applications (such as spam or sockpuppet detection, sentiment analysis, among others), or the emergence of new algorithms (such as deep learning based algorithms).

Researchers have used several approaches based on ML such as Support Vector Machine (SVM) ([3]-[4]), Decision Trees [5], ensembling algorithms and, recently and due to the good results obtained, deep learning based algorithms such as Gated Recurrent Unit (GRU) [6],Long Short-Term Memory (LSTM) [7] or novel approaches that use techniques such as the extended variational inference (EVI) framework to learn the finite inverted Beta-Liouville mixture model (IBLMM) [8]. SVM, ensemble learning or deep learning based models are considered black-boxes, because do not provide a clear interpretation on how the model conclude a result, but, on the contrary, obtain better performance in their evaluation metrics in different contexts and are being widely used. Therefore, XAI is nowadays one of the most relevant research problems [5], [9]-[11].

There are several categories to classify the results of the research work in the XAI field [9]-[10], [12]. The following categorization is proposed in the present work:

1. Model intrinsic approaches. White box models, such classification trees, fall into this category. The model itself is transparent or interpretable, so the user can infer why and how a result was obtained [9], [13].
2. Model specific approaches. These solutions aim to provide a justification based on the algorithm mechanism itself. As Adadi and colleagues stated [10], when a specific type of interpretation is needed, only models that fit that kind of interpretation can be used. These approaches can rely on techniques such as visualization, feature importance or rule extraction on the classification process inside a neural network, among others. For example, extraction of fuzzy rules from a neural network or a SVM [3], [4], the use of a heatmap to interpret a trained SVM model [14], or proposal of a method to decompose the classification decision according to the contribution of the input elements [4].
3. Model independent approaches. There are new techniques aimed to offer an explainable solution from any classifier. Two of the most well-known techniques are: (a) LIME [15], based on the idea of building linear models locally close to the predictions of a black-box model and their variations; and (b) SHAP values [16], which assigns to each variable used by the ML model an additive feature importance value for each prediction according to several desirable properties such as missingness, consistency, and local accuracy. These two approaches do not necessary consider the user in the process of explanation, nor what are the results are going to be use for. They can answer a question "why" or "how" using a post-hoc interpretation, but they do not integrate the users, the data, the model and the context in the process of performing a classification and retrieving an explanation.

Nonetheless, mimetic classifiers, proposed in the first decade of 2000's, do not rely on specific techniques or modifications to the algorithms, but on the workflow defined to classify and obtain a better understanding of the classification results. In essence a machine learning model acts as an oracle that label randomly created examples and then use a second comprehensible model [17], [18], [19]. This solution aims to provide an explainable and resource efficient approach to ensemble algorithms, but its main goal is not to provide a system integrating data, user, models or the context to generate explanations. It is also worth mentioning two recent frameworks: "A framework for explainable text classification in legal document review" [20], which identifies snippets of text that are relevant for the purpose of the review. It is a domain-specific framework which provides explanations based on examples, and it does not explore the relationship between the terms and the topic of the texts. On the other hand, "explAIner" [21] is an interactive and iterative framework based on Visual Analytics (VA) and Interactive Machine Learning (IML). This framework helps understand and refine ML models thus, the main objective of the framework is to build robust models and make them comprehensible for users while building the models. This framework involves three different kind of users but all of them have deep or partial knowledge on ML tasks. Although this framework is conceptually model agnostic, it is implemented using deep learning algorithms and its process is very tied to these kinds of algorithms.

Table I shows a comparative analysis performed over the above mentioned frameworks that generate explanations: Mimetic Classifier, Legal Document Review, and explAIner.

TABLE I. Comparison of Frameworks

Features	Mimetic Classifier	Legal Document Review	explAIner
Explanation Generation Workflow	○	✓	×
Data as key Ingredient	○	×	×
Agnostic Model	○	✓	○
External Knowledge Integration	×	×	×
Involvement of Users	×	×	✓
Explanations in Natural Language	×	✓	×
Explanations based on Term's Relevance	×	×	✓
Interchangeable Components	×	×	○
(*) Leyend:	✓ yes	○ partially	× no

None of the analyzed approaches considers at the same time the features present in Table I, which allow the ingredients involved in a classification task (data, model, users and context) to work together to answer the questions that are relevant for end users and developers to open a black box model (what, how, why). explAIner and mimetic classifiers do not define a workflow to generate explanations, Legal Document Review is not a general purpose natural language explanations based on terms relevance and making data a key ingredient that can integrate external knowledge. None of the frameworks consider data as an ingredient that participates in the process from the beginning both to classify and to obtain explanations. OBOE is based on the idea that the vocabulary that discriminates a document must also be relevant to justify how that document has been classified. Conversely, our framework is designed to integrate external knowledge (Sections III and IV) to provide a general purpose explanation in natural language. Neither explAIner nor Legal Document Review follow the approach we are proposing in this research work to generate explanations, this is, to use natural language based on the relevance of the terminology used in the text. Finally, although explAIner conceptually is a modulable approach, it does not provide general purpose explanations, but to explain the model while building it.

III. Framework Description

We propose a framework called OBOE in which classification of a text can be explained (a) through the early identification of relevant terms in a corpus of documents and (b) from the machine learning techniques used to classify the several documents in the corpus. This justification is internally expressed as symbolic rules and presented in natural language to the user.

All the components defined in OBOE can be customized according to the needs or peculiarities of the classification and explanation tasks. For example, a user that needs to perform a classification over a large corpus can use embeddings to represent the documents and deep learning based algorithms such as LSTM to generate a classification.

In this section we present OBOE at the conceptual level describing its components and the inputs and outputs of each component. In Section IV we present our custom implementation of each component in OBOE to conduct the experiments detailed in Section V.

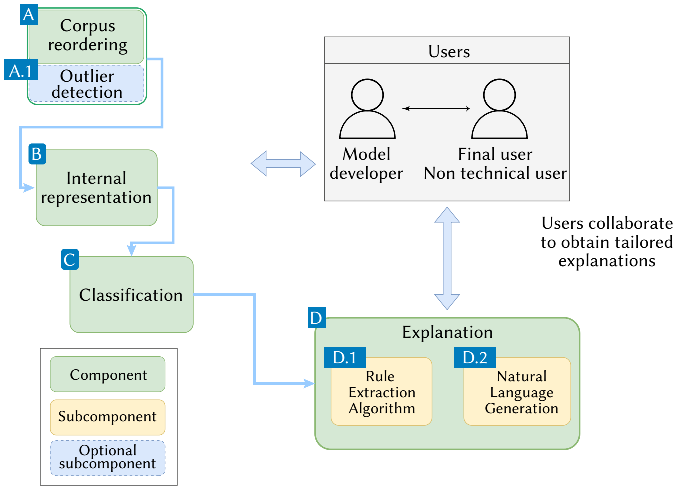


Fig. 1. OBOE components and workflow.

Fig. 1 shows OBOE workflow, where each box is a component of the framework. The components are the following: (A) corpus reordering to early identify relevant terms and documents related to topics, with an (1) optional outlier detection sub-component to create an optional categorization of documents related to the topic we want to discriminate (our target variable), (B) internal representation generation of the corpus, (C) classification algorithm used, and (D) explanation creation by (1) the generation of rules that will be expressed in (2) natural language. OBOE is conceived as a customizable framework since the implementation of each component can be defined by the model developer, which means that the algorithms and techniques are interchangeable.

A. Corpus Reordering

In a traditional classification problem, either binary or multi-class, the corpus is correctly labeled. In others, on the contrary, the corpus of documents is not labeled completely, or is not labeled at all.

In OBOE, data (in our context, a corpus of documents) play a leading role in obtaining an explanation, since it contains the terms that discriminate one class from another. The aim of this component is to identify which terms of the documents that compose the corpus (input of this component) can be useful to explain from the beginning of the process the results of the classification problem.

In this sense, unsupervised techniques such as clustering or topic modeling with any number of clusters or topics can be used to assign new classes to documents and to use the vocabulary identified in such classes to ease the classification and the explanation results. The output is a labeled corpus after applying a technique such as topic modeling or clustering. This output will be later used either by the subcomponent “Outlier detection”, or by the classification algorithm, and also the user (model developer with the collaboration of final users) can extract relevant terms to be used in the explanation component. The participant involved in this component is the model developer, who parametrizes the technique used to reorder the corpus.

1. Outlier Detection

This subcomponent is optional in the designed process of OBOE and can be used, for example, to transform a binary classification into a multi-label classification problem. This component helps to discriminate which documents are related to a given topic, but belong to another topic. For example, a user might be reading documents about Ancient Rome and then find in the corpus a review of the movie ‘Gladiator’. Although this review is going to present words related to documents describing Ancient Rome and is somehow related to Ancient History, there are other words pointing out this is a different kind of text. To that aim, an outlier detection technique, such as Interquartile Range (IQR) or Isolation Forest [22], is applied to the reordered corpus (which acts as an input of this component), to identify documents that are similar to documents of a specific topic. The output of this component is a corpus with a new label assigned to those documents that are similar to other documents in one of the categories that belong to the corpus.

The participants involved in this subcomponent are (a) the model developer, who parametrizes the technique used to reorder the corpus and, (b) the final user who identifies the topic needed to discriminate.

B. Internal Representation

This component translates every document of the reordered corpus (input) to an internal representation (output) to carry out the classification task. In a text classification task, documents are processed (for example, removing special characters and stop words), and transformed into an internal representation that can be managed by a machine learning classifier. Some of the techniques used to preprocess the document are removing stop words and special characters, tokenization, stemming or lemmatizing the texts, among others. Also, common techniques used to generate an internal representation are translating tokens to identifiers a Document Term Matrix (DTM) based on Term Frequency, Term-Frequency Inverse Document representation or those based on embeddings [23].

The participant involved in this component is the model developer that chooses the algorithm accordingly to the problem needs.

C. Classification

This component is a classification algorithm (for example, SVM), and some technique aimed to obtain the variable importance by means of other post-hoc techniques such as LIME or the variable importance identified by any ML framework. The input of this component is the internal representation that will be used by the classification algorithm. The output of this component is a model (the object that represents the algorithm trained) and a representation of the relevance of the words for the classification task.

The participants involved in this component are the model developer and the final user who helps the model developer to parametrize the component and can identify which words can be useful for the explanations.

D. Explanation

This component explains the results obtained during the classification process. To this end, this component is composed of two subcomponents: 1. Rule Extraction Algorithm and 2. Natural Language Generation.

Rules (subcomponent D.1) are contrastive and transparent explanations that can be translated to natural language to exemplify why that text has been classified with a specific class. This eases the collaboration between model developers and final users. Also, rules can manage the importance of the terms in the corpus. This importance of the terms in the documents can be captured in a DTM, regardless of whether an embedding and neural network-based approach had been used for classification or whether knowledge resources are used to enrich the rules or the subsequent natural language explanations.

A natural language explanation (subcomponent D.2) is one of the most straightforward methods to clarify a justification. To generate a natural language explanation a model developer can use a fine tuning approach based on transformers such as T5 [24], or to create and/or to translate the rules into natural language.

To generate the explanations, the explanation component can use as an input the vocabulary identified by the corpus reordering component, the most relevant words identified by the classification framework or another post-hoc technique such as LIME or SHAP [14][15], and the internal representation generated by the internal representation component. This document retrieves a natural language explanation based on rules as an output.

The participants involved in this component are the model developer, who parametrizes algorithms used to obtain the rules and the natural language explanation and the final user who identifies the vocabulary, the size of the rules or the specific text of the explanations, among others.

IV. Framework: Custom Implementation

This section covers the custom implementation made to carry out the experiments described in Section V. As we explain in detail in Section V, we address a binary classification problem and therefore our customized components are conceived to solve this problem.

A. Corpus Reordering: Custom Implementation

As stated in Section III, terms appearing in the corpus (input of this component) are key elements to classify the documents and to explain that classification. In OBOE, these terms help to early discriminate the subject of the text (the class to which the text must be classified, also referred in this paper as "positive class") at the beginning of the process, and to later classify and explain the results.

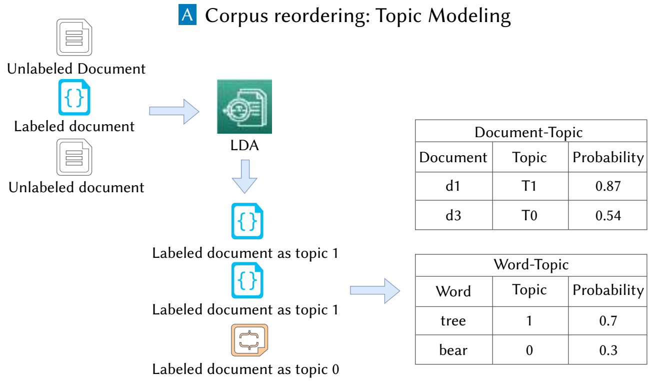
Positive Unlabeled Learning (PUL) is one of the techniques that fits into this scenario. PUL [25]-[27] does not require a fully supervised corpus with positive and negative texts. Instead, PUL uses positive and unlabeled datasets to early discriminate which texts (words) belong to the positive class and their associated probability.

The custom implementation follows a two-step method composed of Topic Modeling with a Latent Dirichlet Allocation (LDA) [28]-[29]. LDA is an unsupervised technique based on Dirichlet probability distribution. Documents are represented as bags of words, from which another one is generated in such a way that each document is a probability distribution of topics, and each topic is a probability distribution of words. In LDA, each document is a mixture of several topics described by a probability distribution that defines how likely each word will appear in each topic.

We apply LDA setting the number of topics equal to 2, as we performed a binary classification of the text. After applying LDA algorithm over the dataset, we obtain:

1. Documents belonging to negative or positive classes identified by LDA, referred as "topic 0" and "topic 1" respectively.
2. The words that likely belong to each topic following the estimated distribution by LDA.

Finally, the output is a corpus of documents labeled as "topic 0" and "topic 1" and their probabilities to belong to the labeled class. Also, we have the probabilities of every term to belong to "topic 1" or "topic 0". Fig. 2 shows a visual example of this component. There are positive and unlabeled documents in a corpus as an input that, after applying the LDA algorithm, are labeled as "topic 0" and "topic 1" with a certain probability. In addition, the different terms appearing in the documents are related to such topics with probability.


A
Corpus reordering: Topic Modeling
Unlabeled Document
Labeled document
Unlabeled document
LDA
Labeled document as topic 1
Labeled document as topic 1
Labeled document as topic 0

Document-Topic		
Document	Topic	Probability
d1	T1	0.87
d3	T0	0.54

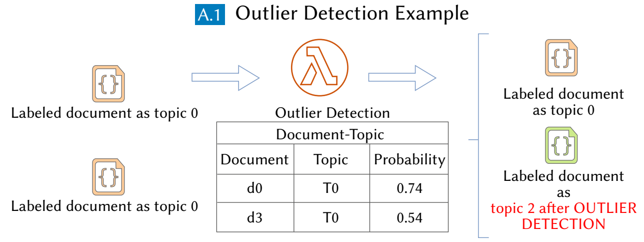
Word-Topic		
Word	Topic	Probability
tree	1	0.7
bear	0	0.3

Fig. 2. Example of Corpus reordering.

1. Outlier Detection Subcomponent: Custom Implementation

We implemented this optional subcomponent in our custom implementation. The input we used was the labeled dataset retrieved by the Topic Modeling main component.

In the context of this work, documents that are related to the positive class might be considered as outliers according to the probability given by LDA of the document belonging to negative class. Therefore, we used the lower bound of an Interquartile range to detect and, therefore, classify the outlier documents.


A.1
Outlier Detection Example
Labeled document as topic 0
Labeled document as topic 0
Outlier Detection
Labeled document as topic 0
Labeled document as topic 2 after OUTLIER DETECTION

Document-Topic		
Document	Topic	Probability
d0	T0	0.74
d3	T0	0.54

Fig. 3. Example of Outlier Detection subcomponent.

The output of this component is a corpus of documents classified into a positive class (‘topic 1’), negative class (‘topic 0’) and positive-related class (‘topic 2’, the outliers). Fig. 3 illustrates an example in which Document 3 (d3), which had a low probability to belong to “topic 0”, is labeled as “topic 2”.

B. Internal Representation Generation: Custom Implementation

This component aims to transform the corpus of documents (input), which is labeled as positive and negative classes and, optionally, with a positive-related class, into an internal representation that can be managed by any classification algorithm

In the context of text classification, a DTM is one of the possible representations that can be used to solve the problem. A DTM is a corpus representation in which rows represent the documents, each column contains a term and cells show a metric about the relative relevance of the term in the document or in the corpus. This metric may be the TF-IDF, calculated as the product of Term Frequency (TF) and the Inverse Document Frequency (IDF):

1. The TF (Term Frequency) is the frequency of the term ‘t’ in a document ‘D’.
2. The IDF (Inverse Document Frequency) is the logarithm of the inverse occurrence rate of the term in the corpus.

The DTM metric used in this module is the inter class dispersion scheme [30], a variation of the DTM explained above, used to enhance the relevance of certain terms with respect to its class.

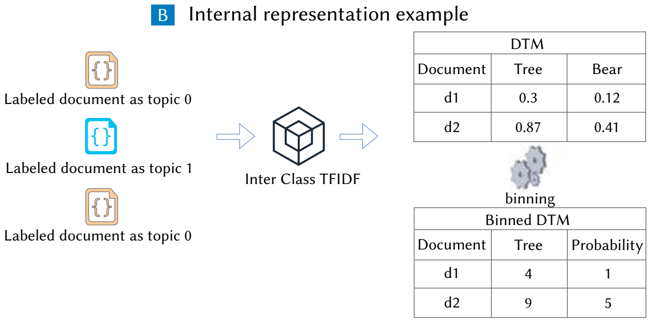
An inter-class dispersion scheme adds a new term to the equation called Dispersion (D(t)). This term of the equation will be low if the term is distributed uniformly among classes, so it helps to identify which terms are ‘good’ for classification. The scheme is described in Equations (1) and (2) [30]:

$$D(t) = \frac{1}{n}\sum_{c=0}^{n}((F(t,c) - \frac{1}{n}(\sum_{c=0}^{n}F(t,c)))^2 \qquad (1)$$

$$Weight(i,j) = TF(i,j) * IDF(i) * D(i) \qquad (2)$$

Equation (1) describes the inter-class dispersion coefficient of the term t. In this equation, “n” is the number of classes and F(t,c) is the number of documents having the term “t” and belonging to the class “c”. Equation (2) represents the weight of the ith term in the jth document.

The output of this component is a DTM. Also, a binning process is applied to obtain a binned DTM. An example is shown in Fig. 4.


B Internal representation example
Labeled document as topic 0
Labeled document as topic 1
Labeled document as topic 0
Inter Class TFIDF
binning

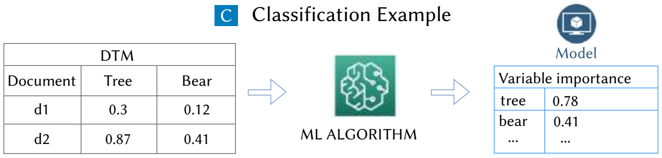
DTM		
Document	Tree	Bear
d1	0.3	0.12
d2	0.87	0.41

Binned DTM		
Document	Tree	Probability
d1	4	1
d2	9	5

Fig. 4. Example of Internal representation.

C. Classification: Custom Implementation

In the implementation we made of this component, we used algorithms based on bagging (the implementation of the Random Forest [31] algorithm by H2O[1], hereinafter DRF) and boosting (XGBoost, hereinafter XGB) [32]. These algorithms used the DTM obtained in the previous component as an input. Once the training is complete, our output is a model that can be used to classify documents and the variable importance identified by H2O. Fig. 5 shows an example of this component.



DTM		
Document	Tree	Bear
d1	0.3	0.12
d2	0.87	0.41

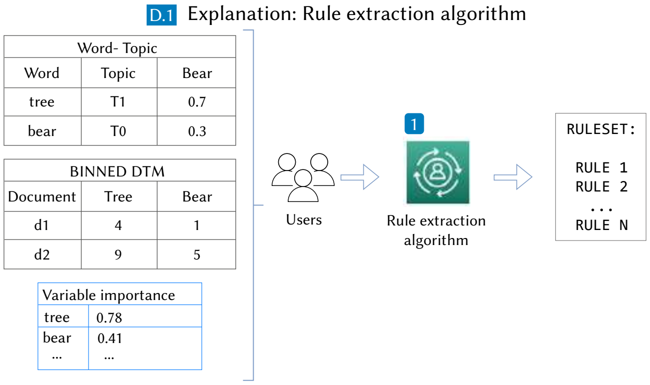
Variable importance	
tree	0.78
bear	0.41
...	...

Fig. 5. Example of Classification.

D. Explanation: Custom Implementation

Our custom implementation of the explanation component comprises two subcomponents: (1) Rule Extraction Algorithm and (2) the Natural Language Generation, which creates a natural language based explanation in Spanish and English. In the custom implementation we made of OBOE, the algorithm generates a rule set using the terms that discriminate a topic to later generate a natural language explanation that uses WordNet[2] to ease the understanding of the explanations adding definitions that facilitate the disambiguation of the term. The rule set with WordNet definitions will finally result in general purpose explanations in natural language according to the needs of final users.

Subsections D.1 and D.2 explain the custom implementation we made of the subcomponents Rule Extraction Algorithm and Natural Language Generation. Fig. 6 and Fig. 7 show an example of the process. The binned DTM in terms of relevance, the variable importance and the probabilities relating words with topics act as input of the component. Then, model developers apply a Rule Extraction Algorithm to obtain a rule set (Fig. 6). This rule set is translated into natural language, adding information of an external knowledge resource (Fig. 7).


D.1 Explanation: Rule extraction algorithm
Users
1
Rule extraction algorithm

Word- Topic		
Word	Topic	Bear
tree	T1	0.7
bear	T0	0.3

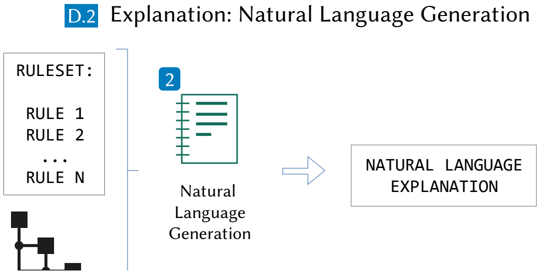
BINNED DTM		
Document	Tree	Bear
d1	4	1
d2	9	5

Variable importance	
tree	0.78
bear	0.41
...	...

```
RULESET:

RULE 1
RULE 2
 ...
RULE N
```

Fig. 6. Example of Explanation: rule set generation.


D.2 Explanation: Natural Language Generation
2
Natural Language Generation

```
RULESET:

RULE 1
RULE 2
 ...
RULE N
```

```
NATURAL LANGUAGE
EXPLANATION
```

Fig. 7. Example of Explanation: natural language explanation.

[1] https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drf.html

[2] https://wordnet.princeton.edu/

1. Rule Extraction Algorithm: Custom Implementation

The input of this subcomponent is the DTM, a vocabulary that can either be specified by the user of this framework (or all the terms of the DTM), as well as other parameters, such as the minimum term relevance to be considered, the length of the rules or the number of rules, among others as presented in this section.

As mentioned in Section I, OBOE is not intended to fit to a specific ML model, but to try to ease the explanation of any possible model. For this reason, in this first subcomponent we generate a rule set. A rule set is a transparent method that can ease the understanding of a classification result and ease the collaboration between model developers and final users. To obtain a set of rules that helps to explain the results of any ML model classification, we implemented a Rule Extraction Algorithm, which is presented as Algorithm 1.

Intuitively, for any ML model, there are variables, that isolated or in combination with others, explain better the target variable. In the context of Natural Language Processing (NLP), these variables are terms whose relevance in the document is defined by a TF-IDF scheme. Instead of using the values of the inter-class TF-IDF weight scheme of the DTM, we used the binned values of each column of the DTM. These values still represent the importance of a term, but discretized in such a way that is easier to map these values to a grade of importance in natural language.

Algorithm 1 uses the binned relevance to generate a rule set with two main steps:

1. Identify the most relevant terms to explain a classification. These terms are chosen according to their information gain with respect to the target.
2. Identify the relevance values of each term in every rule being constructed. This value of relevance is the one that: (i) either minimizes the relation of negative examples vs. positive examples or (ii) increases the number of positive cases with respect to a previously selected value of relevance for that term by a percentage (called Growth Factor).

Both model developers and final users have an important paper to ease the explanations using Algorithm 1, by specifying the input parameters of the algorithm. These parameters define the output of the algorithm that is a set of rules to be translated into natural language. The input parameters are:

1. Vocabulary (V): Algorithm 1 is based on the idea that the vocabulary obtained by the Topic Modeling along with the one obtained as feature importance of the ML model or even the knowledge of the domain of the final user, can explain the results of the classification process.
2. Number of rules (N) and rule length (L): Rule length or coverage might be a drawback in terms of interpretability [3]. To overcome that problem, the algorithm user can specify these parameters to obtain a meaningful rule set.
3. The minimum relevance (min_rel): to enforce the algorithm to use values of a minimum relevance, the user can specify a minimum relevance parameter
4. Growth Factor (GF): this parameter controls the coverage of the rules. With a very large growth factor, the algorithm may result in rules explaining very specific cases (for example, a rule that identifies two or three positive cases with no negative cases), whereas a very small growth factor may include rules that are rather generalist. In our experiments we set up this parameter to 5%.
5. The maximum number of repetitions of terms in the rule set (MR): As the algorithm is based in its first step in the information gain of the term with respect to the target variable, the user can specify the maximum number of repetitions of the term in the rule set to introduce variability.

The Rule Extraction Algorithm (Algorithm 1) is as follows:

```
Algorithm 1: Rule Extraction Algorithm
I:  Document Term Matrix (DTM), number of rules (N), length
    of rules (L), max repetition per term (MR), vocabulary (V), target
    variable (T), growth_factor (gf), min_relevance (min_rel)
O:  Set of rules
begin  Initialize variables:
1  if V is Null then V ⟵ get column names of DTM
2  positive_examples ⟵ Number of rows of DTM with T = 1
3  number_of_rules=1; length_of_rules=1; rule_set= empty set;
   DTMoriginal = DTM;
4  while positive_examples>0 and number_of_rules <=N:
5    to_exclude ⟵ empty list
6    while length_of_rules<=L:
7      attr ⟵ getVariableWithMaxInfoGain (V,T,DTM,MR,to_exclude)
8      val ⟵ getValueOf Term(attr,DTM,GF,min_rel)
9      Coverage ⟵ getPositivesRate(attr,val,DTMoriginal)
10     to_exclude.add(attr)
11     r.add(attr,val,coverage)
12     length_of_rules⟵ length_of_rules + 1
13     positive_examples, DTM ⟵ filterDTM(DTM, r )
14   rule_set.add(r)
15   number_of_rules ⟵ number_of_rules + 1
End  return rule_set
```

Steps 7 and 8 are the two main steps of Algorithm 1: get attribute that maximizes the information gain with respect to the target, and for the selected term, get the value (that must be greater than min_rel) that minimizes negative vs. positive cases or that increase the positive cases in a Growth Factor. Therefore, the algorithm calculates the number of positive examples against the total to get the coverage of the rule in the Step 9.

The algorithm will stop when the number of rules is equal to the number of rules specified by the user or when there are not more positive examples to cover.

After creating the rule set, a consolidation step is applied. This step identifies rules with the same antecedents and different values to perform a simplification of the rule set, and obtaining the coverage of the new rule. Fig. 8 depicts an example of the left hand side of several rules and its consolidated form.

```
book == 4 and read == 8
book == 4 and read == 7
book == 4 and read == 9
```

(a)

```
book == 4 and ((read>=7 and read<=9))
```

(b)

Fig. 8. Part (a) shows a bunch of left-hand side rules; part (b) is the left hand side of the new rule consolidated.

Finally, a rule is created with in the form "`if LHS then coverage is 0.X`", in order to improve its understandability, in the Natural Language Generation component, the rules are translated into the form: "`if LHS then is a '<name of the class>'`". In this point, information about the coverage and the rate of negative over positive examples that the ruleset covers is added for the model developer's knowledge.

This Rule Extraction Algorithm can be thought as a simplified variant of IREP [33], but differs in the use of Mutual Information and in the fact of that it is not thought to maximize any accuracy metric, but to ease the understanding of any classification.

The result of this subcomponent (output) is a set of rules that will be translated into natural language by the next subcomponent.

2. Natural Language Generation: Custom Implementation

The set of rules obtained by the Rule Extraction Algorithm component (input) will be used to generate a natural language translation from each rule (output). The subsequent translation into natural language is structured in such a way that the results of the classification can be understood by any user, regardless of his/her previous knowledge of the domain.

To generate the natural language explanation from the rules, we used a Context-free Grammar (CFG). The values indicating the relevance of the term in the text and corpus are translated into natural language in terms of importance, as shown in Table II.

TABLE II. Scale of Importance

Relevance	NL Translation – Spanish	NL Translation - English
0-2	Muy poco importante	Very unimportant
3-4	Poco importante	Unimportant
5-6	De Importancia media	Of medium importance
7-8	Importante	Important
9-10	Muy importante	Very important

The rules parsed using the context-free grammar create a natural language explanation that is completed using WordNet with the information retrieved by the definitions of the concepts that are contained in the rules. Appendix I shows an example of explanation in Spanish and the same one translated into English.

V. Evaluation of the Custom Implementation of OBOE

We used three corpora to perform the experiments on the custom implementation of OBOE. These three corpora were all processed using the same pipeline by removing stop-words, special characters, and lemmatizing the words:

1. Amazon review corpus [34]. It contains the review of near 6 million objects (books, cell phones, etc.), divided into several categories such as Books, Electronics, Cell Phones & Accessories, Office Products, or Home & Kitchen, among others. We randomly selected 25.000 Books reviews and 25.000 reviews of other categories, labelled as positive class the 25.000 Books reviews.
2. Reuters dataset [35] included in the NLTK package[3]. It contains 10.788 documents from the Reuters Financial Newswire Service divided in 90 categories, although a document can belong to more than one category. The documents having the class 'acq' are the positive class.
3. 20 Newsgroup dataset [36], included in the scikit-learn package4. It contains 18000 documents divided into 20 categories. The documents belonging to category number '3' ('comp.sys.ibm.pc.hardware') are the positive class.

We present in this section the results obtained in the Amazon corpus, whereas the results obtained with the other two corpora are detailed in Appendices II and III. This section presents the results obtained in the experiments to evaluate whether LDA can discriminate the two possible categories of each target variable (one per corpus), and the classification performed by XGB and DRF. The specific algorithms and techniques can be changed by others if they fit the workflow defined by OBOE. The implementation employs Python programming language, scikit-learn5, spacy6, NLTK , and H2O.

A. Evaluation of Corpus Reordering

We performed an evaluation on the results obtained by LDA, comparing the topic assigned by LDA with the actual topic of the documents. We used the probability distribution for the documents belonging to "topic 0" or "topic 1" as the predictor variable, and then we trained a XGB model using the real target as variable to predict. The hyperparameters used in the XGB models are based on the results retrieved by Bayesian Optimization [37].

Appendix II contains the performance, the hyperparameters and the confusion matrix obtained in the three corpora.

The performance obtained in terms of Area Under Curve (AUC) and Kappa index were w 0.89 and 0.72, respectively, in the Amazon corpus. In addition the error rate of the positive class was 0.07 and the error rate of the negative class, 0.2.

B. Evaluation of Classification Models

We conducted two experiments to evaluate the classification component: the first one evaluates the assignment of topics to each document, without outlier detection subcomponent, and the second one uses the reordered corpus generated by the outlier detection subcomponent that creates a third class with documents that might be related to the positive class. We used XGB and RF to classify the documents and Bayesian Optimization to find the hyperparameters used in each of the trained models. Appendix III contains the tables detailing the performance and confusion matrix.

Regarding the classification experiment without outlier detection, XGB and RF obtained a performance of 0.99 and 0.964, respectively.

In the case of the results of classification with outlier detection, we used an IQR [37] approach over the probability distribution, with an IQR of 1.5. To perform the search of hyperparameters, we used hyperopt[7] package, letting the library to choose the best algorithm to perform the optimization. We used the LogLoss error as the cost function. The LogLoss of the classification performed by DRF and XGB was 0.32 and 0.36, respectively. It is worth mentioning that only the Reuters dataset obtained an error rate lower than 0.5 in class 2 (documents related to main topic). In most of the cases, the classifier erroneously classified these documents as positive class. Nonetheless the error rate obtained in Reuters suggests that the approach seems to be valid (see Appendix III).

C. Rule Extraction Algorithm

For each dataset, we obtained two set of rules with the Rule Extraction Algorithm defined by Algorithm 1. In the first one, the maximum number of rules must be 10. Besides, in the second one, the maximum repetitions per term is three and the vocabulary was also provided.

Specifying the vocabulary and the maximum repetitions per term permits the model developer, interacting with the final user, to obtain a compact set of rules that ease the further understanding of the results achieved during the classification. Therefore, the framework involves the user in all the process through an analysis of the most relevant terms identified by the components. Also, users might include vocabulary if they have previous knowledge of the domain. Fig. 9 presents a sample of rules obtained for the

[3] https://www.nltk.org

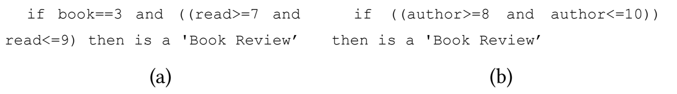
[4] https://scikit-learn.org/stable

[5] https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html

[6] https://spacy.io

[7] https://github.com/hyperopt/hyperopt

Amazon corpus. Fig. 9a shows an example specifying the maximum number of rules, whereas Fig. 9b shows an example specifying the vocabulary:{'word','book','romance','story','author','character','read'}.



```
  if book==3 and ((read>=7 and      if ((author>=8 and author<=10))
read<=9) then is a 'Book Review'    then is a 'Book Review'
```

(a) (b)

Fig.9. Part (a) is shown a bunch of left hand side rules; part (b) is the left hand side of the new rule consolidated.

It is also worth mentioning that two different classifiers can obtain a variable importance whose similarity may be influenced by factors such as the inner logic and the parameterization of the algorithms, among others. This may affect the vocabulary that the model developer uses to generate the rules and the natural language explanation. Also, the model developer can consider the use of other techniques to obtain the 'relevance' of the term to the prediction of the target variable. As a matter of example, Appendix IV contains 4 figures depicting the variable importance obtained by DRF and XGB algorithms both in the Amazon Reviews and Reuters corpora. These figures show that the variable importance obtained by both algorithms in the Amazon Review corpus is more similar than the obtained by the algorithms in the 20 Newsgroup corpus. For this reason, the collaboration of model developers and final users is important to obtain meaningful explanations.

VI. User-based Evaluation on Explanations

To gather subjective empirical data about the explanations provided by OBOE, a questionnaire was implemented as a Google Form[8] and launched via different mailing lists, posts in LinkedIn, and tweets to obtain responses from general population. We collected 38 responses and removed two of them from the analysis because of the inappropriate responses about the professional affiliation.

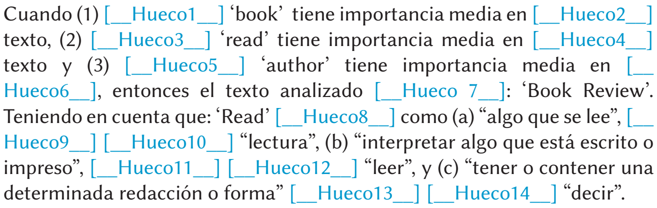
This questionnaire was divided into three subdivisions: the first one focused on questions about the level of comprehensibility (ease of understanding) and legibility (ease of reading) of the explanations generated by OBOE, the second one devoted to get suggestions and recommendations, and the final part related to demographic data.

More specifically, the first subdivision of the questionnaire included two parts: (a) one related to the level of comprehensibility of explanations provided by OBOE, and (b) another one related to the level of legibility of such explanations. In the case of (a), the questionnaire included three questions based on Cloze Test [39]. The idea behind these types of questions was to analyze whether the whole text, with removed elements, is understood, and thus, the removed elements are filled in with logical language items. In the case of (b), the questionnaire included three questions based on the idea of binary forced choice [40]: "humans are presented with pairs of explanations and must choose the one that they find of higher quality (basic face-validity test made quantitative)." Each question is divided in several sub questions to evaluate key parts of the explanation provided by OBOE, such as the appropriate selection of the terms and the alternatives to the redaction provided by OBOE. The total amount of sub questions is 16, so the total amount of responses analyzed in the first part of the questionnaire are 576. With this type of questions, we can conclude whether the explanation pattern is comprehensible and whether there are patters which are preferred for an explanation.

According to the questionnaire, the 65.8% of the respondents were in an age range between 31 and 45 years old and have a university degree, whereas the 25.8% of the respondents are PhD. The profession of the 71.1% of the respondents is related to computer science.

A. Comprehensibility

The three questions of this questionnaire part were related to three different explanations: first and second questions were related to a short and a large explanation respectively, and the third question was related to que adequateness of the term 'importance' versus 'relevance'. The total amount of sub-questions in the comprehensibility section is 12. Fig.10 shows an example of question of this part:



Cuando (1) [__Hueco1__] 'book' tiene importancia media en [__Hueco2__] texto, (2) [__Hueco3__] 'read' tiene importancia media en [__Hueco4__] texto y (3) [__Hueco5__] 'author' tiene importancia media en [__Hueco6__], entonces el texto analizado [__Hueco 7__]: 'Book Review'. Teniendo en cuenta que: 'Read' [__Hueco8__] como (a) "algo que se lee", [__Hueco9__] [__Hueco10__] "lectura", (b) "interpretar algo que está escrito o impreso", [__Hueco11__] [__Hueco12__] "leer", y (c) "tener o contener una determinada redacción o forma" [__Hueco13__] [__Hueco14__] "decir".

Fig. 10. Cloze test question from the Comprehensibility part.

The 6.9% of the responses gathered in this section are incorrect. These incorrect answers were written by the respondents, without being an option in the Cloze's tests.

The rest of the responses gathered are as follows:

- The 38,2% correspond to the actual explanations provided by OBOE.
- The 41,4% correspond to an alternative that it is synonym of the explanations provided.
- 13,4% are valid responses provided by the respondents.

There are several aspects remarkable in this section:

1. The explanations provided by OBOE do not exactly translate a rule of the form "if - then". In this sense, in the questions where the users have to choose between the alternatives: "Cuando" (whenever), "Si" (if) or an alternative response provided by the user, the majority of the responses retrieved correspond to the option "Cuando", which is the same as OBOE provides. On the contrary, in the subquestion where the respondents have to choose between the wording provided by OBOE "podría tratar de" (may be related to) or the alternative "se clasificaría como" (would be classified as), the majority of the respondents preferred this second option which can be considered more technical from a computer science point of view. Analogously when we asked the respondents to choose between "acepción" (meaning or connotation) or "sentido" (sense) the first one was the most voted, although is less technical from a computer science perspective
2. In the same way, there are two sub questions where the respondents must choose between "importance" or "relevance". In one of them, the majority option selected was "importance" whereas in the second one was "relevance".

B. Legibility

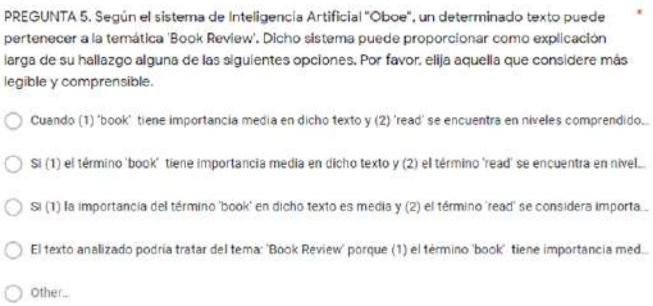
There are four sub-questions in this questionnaire part aimed to analyze different form of redactions. Fig. 11 shows an example of these sub-questions.

The first three sub questions analyze the form of presenting the redaction according to the categorization of a text. The several redactions provided can be divided into two alternatives:

1. The alternative in the form 'if – then' and its derivatives. As said before, the explanation provided by OBOE falls into this group, although it is not exactly translated as an "if - then" question.
2. Another alternative in a less structured language from the point of view of explaining or translating a rule, where (a) the possible classifications of the text goes first and (b) the causes are in the last part of the phrase, for example:

[8] https://forms.gle/jsGC5DjfzWveVnz69

> *"El texto analizado podría tratar de: Book Review porque el término "read" es importante en dicho texto" that can be translated into English as "The text analyzed could be related to the subject: Book Review because the term "read" is important in that text"*



PREGUNTA 5. Según el sistema de Inteligencia Artificial "Oboe", un determinado texto puede pertenecer a la temática 'Book Review'. Dicho sistema puede proporcionar como explicación larga de su hallazgo alguna de las siguientes opciones. Por favor, elija aquella que considere más legible y comprensible. *

○ Cuando (1) 'book' tiene importancia media en dicho texto y (2) 'read' se encuentra en niveles comprendido...

○ Si (1) el término 'book' tiene importancia media en dicho texto y (2) el término 'read' se encuentra en nivel...

○ Si (1) la importancia del término 'book' en dicho texto es media y (2) el término 'read' se considera importa...

○ El texto analizado podría tratar del tema: 'Book Review' porque (1) el término 'book' tiene importancia med...

○ Other...

Fig. 11. An example question from Legibility part.

The 50% of the users chose an 'if-then' related alternative in the three sub questions. The 23,1% selected the alternative provided by OBOE and the 26.9% other alternatives using the prefix "Si" (if) instead of "Cuando" (whenever), what seems somehow contradictory with the results obtained in the comprehensibility part.

The last sub question analyzed the redaction of the definition of the terms. In this case there are five possible options:

1. The one provided by OBOE: Possible definition of <terms>: according to the sense <form>: definition
2. Definition of <term> according to the sense: <form>: <definition>
3. Definition of <term> according to the connotation: <form>: <definition>
4. <term> is defined as <definition> according to the sense <form>
5. <term> is defined as <definition> according to the connotation <form>

The explanation provided by OBOE in this case just obtained one vote. By contrast, most of the respondents chose the options where the term is introduced in the beginning (4 and 5) with 9 and 12 votes, respectively which can be considered as a more informal way to present the explanation.

Nonetheless, from the responses retrieved in this part, it is difficult to conclude whether users prefer an explanation presented in a more informal way, or not. The first three questions show a balance between the options provided in a more structured way and the more informal one. Even the sub question number four of this section cannot be considered unformal, but not as structured as the other options.

Also, analyzing these results with those obtained in the comprehensibility part, we can conclude that users do not prefer terminology related to plausibility, such as "podría tratar de" (may be related to) or "posible definición de" (possible definition of).

VII. Discussion

The results of the custom implementation of OBOE show that the proposed approach achieves comprehensible explanations of the classification process. The ingredients involved in the process of classification and generation of explanations, i.e. data, model, users, play an active role and become relevant to generate explanations. This is a novel approach that none of the analyzed frameworks uses., Furthermore, OBOE aims to define a workflow of interchangeable and customizable components to provide explanations in natural language that can be completed with external knowledge resources. Moreover, OBOE relies on the idea that the same terms which discriminate a document must be used to explain the classification of such a document. This approach eases the use of well known classification techniques and also allows the use of newer ones, which is crucial to allow a tailored solution to the context and user needs. Table III, summarizes our proposal in comparison to other frameworks.

TABLE III. Comparison of Frameworks

Features	OBOE	Mimetic Classifier	Legal Document Review	explAIner
Model	Independent	Intrinsic: needs a White Box classifier	Independent	Deep Learning
Explanations	Rules and Natural Language	No	Examples	Visual Analytics
Explanation Workflow (EW) /Workflow/ Tool (T)	EW	W	T	T
User Involvement	Model Developer and Final Users have the role of orchestrating the workflow	Model Developer	Model Developer	Model Developer
How data is used	Defines model vocabulary and explanations	As usual	As usual	As usual
Relevance of terminology	Yes	No	No	Yes
Customization	Yes	No	No	Partially
Can integrate external knowledge	Yes	No	No	No

While OBOE defines a customizable, model-independent workflow to obtain natural language explanations, none of the other frameworks analyzed fit into this framework. Besides, data and users become relevant during all the process:

- Vocabulary can be defined from the beginning and is based on the relevance of the terms with respect to the class you are trying to predict and explain. As a consequence, data is not used 'as usual', this is, just as the input of the classification process.
- Users are orchestrators of the workflow, specifying the parameters that lead to an ulterior explanation: with the importance of the terms obtained by the ML model and LDA, the user indicates which are the main terms that help explain the subject of the text. Then, both model developers and final users generate a rule set that ease the interpretation and comprehension of the results by any kind of user. So, the user (model developers and final users) is controlling how the explanations need to be obtained, which is an aspect related to the context.
- It is worth noting that Mimetic Classifier defines a workflow, but this workflow is not intended to generate explanations but to ease the understanding of the classification results using a white box classifier.

OBOE is flexible to use any ML model or optimization technique to better classify the documents, and this feature can lead to different results along the process; other frameworks such as Mimetic Classifiers or explAIner are tied to the machine learning technique used. In our custom implementation we used PUL, so we assumed that we did not know the composition in terms of classes of every document of the corpus. Also, we could try the use of Topic Modeling instead of clustering which is not a common way to approach a PUL problem.

The use of specific techniques can lead to different results. In this sense, the variable importance obtained by a ML algorithm can vary, so the user can obtain a different vocabulary to generate the explanations. This feature is also crucial as not every algorithm does fit any problem. Also, different techniques can be used to improve the generation of rules, or to select the most relevant variables in the light of the results obtained in the model (LIME, SHAP, among others). This flexibility can also help users (model developers and final users) to adapt the explanation to the context, by integrating external knowledge or even using the visualizations to complete the natural language explanations.

Finally, the customization feature it is also crucial from a comprehensibility perspective. Results show in section VI that it is not even clear which language a user might prefer. The adaptability to any context and user allows the framework to adapt the language of the explanations to the context of final users and also use knowledge resources such as vocabularies or ontologies.

VIII. Conclusion

This paper presents OBOE, a text classification framework, which aims to provide meaningful explanations. Data, model, explanations and users are ingredients involved in the classification process that play an active role along the process of classification and generation of explanations. OBOE defines various components that can be customized according to the specific context, users and needs, both model developers and final users.

In order to show the customizable feature of OBOE, a specific implementation is presented based on LDA for corpus reordering, IQR for Outlier Detection, an inter-class dispersion scheme for DTM creation, XGB and DRF for classification and a custom Rule Extraction Algorithm and Context-free Grammar to generate general purpose natural language explanations.

We have performed three validations for the implementations we made of the components of OBOE using Amazon, Reuters and 20 Newsgroup corpora: (i) corpus reordering evaluation, (ii) the classification evaluation and the (iii) explanations evaluation. As our customization is based on PUL, the first validation shows whether the topics assigned by LDA algorithm match the actual ones; while the second one evaluates the classification results.

The corpus reordering and classification evaluation achieved an AUC of 0.89 and 0.9, and a Kappa Index of 0.72 and 0.95 in the Amazon corpus. The error rate of the positive class was 7.1%.

In case of the classification component, we performed two different evaluations: with and without outlier detection. We used XGB and DRF algorithms to perform the classification of the reordered corpora. The results achieved in these experiments show that the algorithms discriminate the reordered corpora. When considering the results obtained after Outlier Detection, in the case of Reuters dataset, we obtained and error rate in class 2 below than 0.5. This can be due to several factors, such as the need to use specific techniques for unbalanced multi class classification. Nevertheless, the results achieved suggest that this component is valid.

We also performed a user-based evaluation with the goal of determining whether the explanations provided by OBOE are comprehensible and legible. From the results obtained, we can conclude that the explanations generated by the custom implementation of OBOE are comprehensible, although there is not a clear preference between a more technical or informal language. In the same way, there is not a clear preference between a structured or more informal explanation when presenting the relevance of the terms, there was a tie between a "if-else" based explanation structure and a more informal one. Nevertheless, most of the survey respondents preferred a more informal choice when defining terminology.

Our current and future work is aimed to integrate semantic models in the explanations provided, using linguistic knowledge resources to perform knowledge-based translations and to generate explanations based on the familiarity of the user with a specific domain.

Appendix I: Example of OBOE explanation

A. Example of Explanation in Spanish

Explicación generada de uno de los casos encontrados donde un texto puede tratar de ‘Book Review’

En el contexto de encontrar y justificar la temática de un texto, hemos podido deducir que cuando

(1) character se encuentra en niveles comprendidos entre tiene importancia media y es importante en dicho texto o (2) character es muy importante en dicho texto, entonces el texto analizado podría tratar del tema: ‘Book Review’

Algunas definiciones de los términos arriba expuestos:

Mostrando 3 posibles definiciones para el término: <character>

0. Posible definición para <character>, de acuerdo con el sentido: calidad. Definición: una propiedad característica que define la aparente naturaleza individual de algo

1. Posible definición para <character>, de acuerdo con el sentido: característica. Definición: una propiedad característica que define la aparente naturaleza individual de algo

2. Posible definición para <character>, de acuerdo con el sentido: carácter. Definición: el complejo inherente de atributos que determina las acciones y reacciones morales y éticas de una persona

B. Example of Explanation in English

Explanation generated from one of the cases found where a text can be talking about the subject ‘Book Review’

In the context of finding and justifying the theme of a text, we have been able to deduce that whenever

(1) *character* is at levels between being of medium importance and important in that text or (2) *character* is very important in that text, then the text analyzed could be related to the topic: ‘Book Review’

Some definitions of the above terms:

Showing 3 possible definitions for the term: <*character*>

0. Definition of <*character*>, in compliance with the semantic meaning: quality. Definition: a characteristic property that defines the apparent individual nature of something

1. Definition of *‘character’*, in compliance with the semantic meaning: characteristic. Definition: a characteristic property that defines the apparent individual nature of something

2. Definition of <*character*>, in compliance with the semantic meaning: *character* Definition: the inherent complex of attributes that determines a person’s moral and ethical actions and reactions

Appendix II: Corpus Reordering Performance

A. Performance and Hyperparameters

Corpus	AUC	KAPPA	Hyperparameters	Under sampling
Amazon	0.89	0.72	max_depth: 39, ntrees: 393, min_rows:5, eta:0.6331, learn_rate:0.418, sample_rate:0.2508, colsample_bytree:0.633, reg_lambda:0.86, reg_alpha:0.062	No
Reuters	0.90	0.66	max_depth: 4, ntrees: 400, min_rows: 2, eta: 1, learn_rate: 0.01, sample_rate: 0.5, colsample_bytree: 0.2, reg_lambda: 0, ‘reg_alpha’: 1	In Reuters and 20 Newsgroup corpus, we randomly undersampled the majority class in 3.000 and 6.000 documents, respectively.
20 Newsgroup	0.78	0.13	max_depth: 39, ntrees: 115, min_rows:4, eta:0.1563, learn_rate:0.1745, sample_rate: 0.2589, colsample_bytree:0.5599, reg_lambda:0.924, reg_alpha:0.7721	

B. Confusion Matrix

Corpus	Predicted Class	Actual Class 0	Actual Class 1	Error rate
Amazon	0	4929	1253	0.2027
	1	453	5904	0.0713
Reuters	0	608	160	0.208
	1	68	520	0.115
20 Newsgroup	0	3439	1057	0.23
	1	84	142	0.37

Confusion matrix presented above shows that XGB models predicted in a similar way the negative class in every corpus, with an error rate close to 0.2, and quite well the positive class, with error rates of 7.1% and 11.5%, respectively. The results achieved with 20 Newsgroup corpus point out an error rate of 37% for the positive class, suggesting that the vocabulary of the chosen class is not as specific as in the other corpora.

Appendix III: Classification Performance

A. Performance and Hyperparameters Without Outlier Detection

Corpus	Algorithm	AUC	KAPPA	Hyperparameters
Amazon	DRF	0.964	0.82	max_depth:29, min_rows:15, ntrees:290, sample_rate:0.46
	XGB	0.99	0.95	colsample_bytree: 0.683, eta: 0.37, learn_rate: 0.32, max_depth: 25, ‘min_rows’: 31, ntrees: 365, reg_alpha: 0.5, reg_lambda 0.86, sample_rate: 0.44
Reuters	DRF	0.99	0.96	max_depth:30, min_rows:7, ntrees:191, sample_rate:0.23
	XGB	0.99	0.95	colsample_bytree: 0.655, ‘eta’: 0.776, learn_rate: 0.88, max_depth: 25, min_rows: 3, ntrees: 298, reg_alpha: 0.24, reg_lambda: 0.7, sample_rate: 0.344
20 Newsgroup	DRF	0.87	0.53	max_depth:25, min_rows:2, ntrees:141, sample_rate:0.417
	XGB	0.88	0.57	colsample_bytree: 0.655, eta: 0.776, learn_rate: 0.88, max_depth: 25, min_rows: 3, ntrees: 298, reg_alpha: 0.24, reg_lambda: 0.7, sample_rate: 0.344

B. Performance and Hyperparameters With Outlier Detection

Corpus	Algorithm	LogLoss	Hyperparameters
Amazon	DRF	0.32	max_depth :29, min_rows :15, ntrees :290, sample_rate:0.46
	XGB	0.36	colsample_bytree: 0.683, eta: 0.37, learn_rate: 0.32, max_depth: 25, min_rows: 31, ntrees: 365, reg_alpha: 0.5, reg_lambda: 0.86, sample_rate: 0.44
Reuters	DRF	0.54	max_depth :30, min_rows :7, ntrees :191, sample_rate :0.23
	XGB	0.83	colsample_bytree: 0.31, eta: 0.68, learn_rate: 0.99, max_depth: 38, min_rows: 3, ntrees: 352, reg_alpha: 0.205, reg_lambda: 0.82, sample_rate: 0.5
20 Newsgroup	DRF	0.1	max_depth :25, min_rows :2, ntrees :141, sample_rate: 0.41
	XGB	0.12	colsample_bytree: 0.655, eta: 0.776, learn_rate: 0.88, max_depth: 25, min_rows: 3, ntrees: 298, reg_alpha: 0.24, reg_lambda: 0.7, sample_rate: 0.34

C. Confusion Matrix With Outlier Detection Using DRF

Corpus	Predicted Class *	Actual Class 0	Actual Class 1	Actual Class 2	Error rate
Amazon	0	5925	351	0	0.062
	1	571	598	9	0.082
	2	219	266	0	1
Reuters	0	1916	3	0	0.0015
	1	21	717	0	0.02
	2	54	21	0	1
20 Newsgroup	0	1129	592	7	0.34
	1	207	2646	0	0.07
	2	19	130	9	0.94

*Class 2 corresponds is the assigned to might be related to positive class

D. Confusion Matrix With Outlier Detection Using XGB

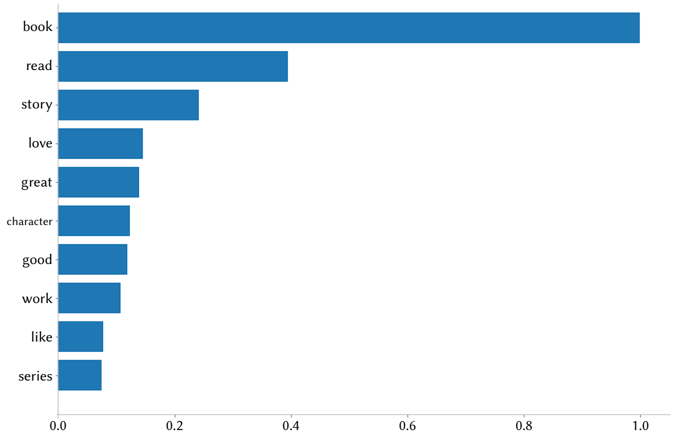
Corpus	Predicted Class *	Actual Class 0	Actual Class 1	Actual Class 2	Error rate
Amazon	0	5311	320	15	0.059
	1	544	5894	31	0.088
	2	228	211	46	0.9
Reuters	0	1906	7	6	0.006
	1	11	716	11	0.029
	2	17	17	41	0.45
20 Newsgroup	0	1253	448	27	0.27
	1	406	2419	28	0.15
	2	30	99	29	0.81

*Class 2 corresponds is the assigned to might be related to positive class

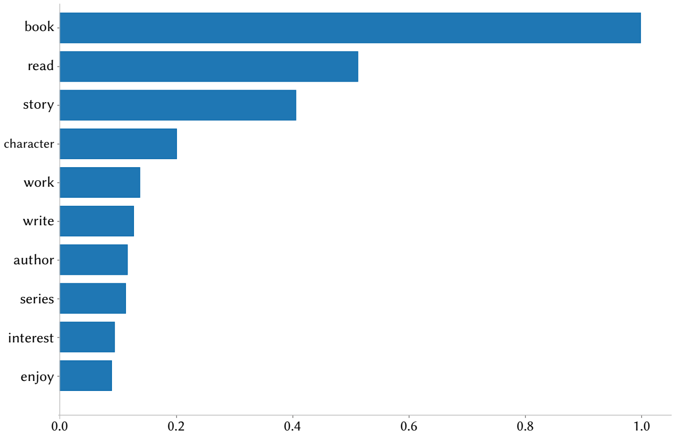
Appendix IV: Variable importance comparison

A. Amazon Review Dataset

1. Variable Importance Described by XGB

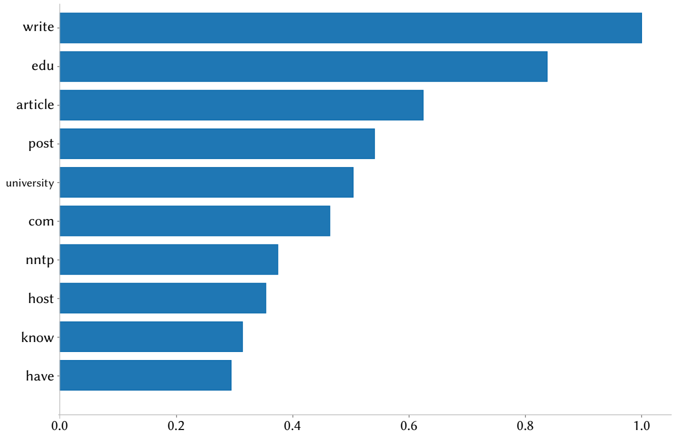


2. Variable Importance Described by DRF

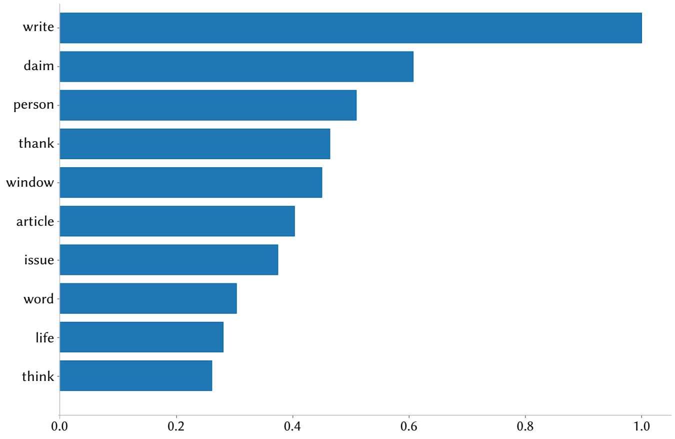


B. 20 Newsgroup

1. Variable Importance Described by XGB



2. Variable Importance Described by DRF



Acknowledgments

This research is supported by the project MiRed (Microrrelato hipermedial español e hispanoamericano (2000-2020) "Elaboración de un repositorio semántico y otros desafíos en la red" (RTI2018-

094725-B-I00))[9]. We would like to thank Daniel Manrique Gamo, Phd., Associate Professor at Universidad Politécnica de Madrid, for his insightful comments, and Tatiana Eman, an English teacher specialized in Applied Linguistics, who provided very useful comments for translating Spanish explanations to English.

References

[1] F. Lecue, “On The Role of Knowledge Graphs in Explainable AI | www.semantic-web-journal.net,” Semantic Web Journal, p. 9, 2018.

[2] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, “Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers,” arXiv:1801.06889 [cs, stat], May 2018, Accessed: Nov. 07, 2020. [Online]. Available: http://arxiv.org/abs/1801.06889.

[3] L. M. Brasil, F. M. de Azevedo, and R. Moraes, “FUZZYRULEXT: extraction technique of if/then rules for fuzzy neural nets,” in Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No.00CH37143), Jul. 2000, vol. 2, pp. 1271–1274 vol.2. doi: 10.1109/IEMBS.2000.897967.

[4] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, “Explaining nonlinear classification decisions with deep Taylor decomposition,” Pattern Recognition, vol. 65, pp. 211–222, May 2017, doi: 10.1016/j.patcog.2016.11.008.

[5] H. Zhang, S. Nakadai, and K. Fukumizu, “From Black-Box to White-Box: Interpretable Learning with Kernel Machines,” in Machine Learning and Data Mining in Pattern Recognition, NY, 2018, pp. 213–227. [Online]. Available: https://www.springer.com/gp/book/9783319961323.

[6] K. Cho et al., “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” arXiv:1406.1078 [cs, stat], Sep. 2014, Accessed: Jul. 10, 2021. [Online]. Available: http://arxiv.org/abs/1406.1078.

[7] S. Hochreiter and J. Schmidhuber, “Long Short-term Memory,” Neural computation, vol. 9, pp. 1735–80, Dec. 1997, doi: 10.1162/neco.1997.9.8.1735.

[8] Y. Ling, W. Guan, Q. Ruan, H. Song, and Y. Lai Y. “Variational Learning for the Inverted Beta-Liouville Mixture Model and Its Application to Text Categorization”. International Journal of Interactive Multimedia and Artificial Intelligence, vol. 7, no.5, pp. 76-84, Sept. 2022.

[9] A. B. Arrieta et al., “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI,” arXiv:1910.10045 [cs], Dec. 2019, Accessed: Feb. 13, 2020. [Online]. Available: http://arxiv.org/abs/1910.10045.

[10] A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” IEEE Access, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[11] D. Gunning, “Explainable Artificial Intelligence (XAI),” Tech. rep. Defense Advanced Research Projects Agency (DARPA), 2017.

[12] Z. C. Lipton, “The Mythos of Model Interpretability,” arXiv:1606.03490 [cs, stat], Mar. 2017, Accessed: Feb. 11, 2020. [Online]. Available: http://arxiv.org/abs/1606.03490.

[13] L. S. Prasanthi and R. K. Kumar, “ID3 and Its Applications in Generation of Decision Trees across Various Domains- Survey”. In (IJCSIT) International Journal of Computer Science and Information Technologies vol. 6, p. 5, 2015. [Online]. Available: http://ijcsit.com/docs/Volume%206/vol6issue06/ijcsit20150606109.pdf.

[14] L. Rosenbaum, G. Hinselmann, A. Jahn, and A. Zell, “Interpreting linear support vector machine models with heat map molecule coloring,” Journal of Cheminformatics, vol. 3, no. 1, p. 11, Mar. 2011, doi: 10.1186/1758-2946-3-11.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16, San Francisco, California, USA, 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[16] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” presented at the NIPS, Dec. 2017. [Online]. Available: https://www.researchgate.net/profile/Scott_Lundberg2/publication/317062430_A_Unified_Approach_to_Interpreting_Model_Predictions/links/5a18eb21a6fdcc50ade7ed19/A-Unified-Approach-to-Interpreting-Model-Predictions.pdf.

[17] P. Domingos, “Knowledge discovery via multiple models,” Intelligent Data Analysis, vol. 2, no. 1, pp. 187–202, Jan. 1998, doi: 10.1016/S1088-467X(98)00023-7.

[18] R. Blanco-Vega, J. Hernández-Orallo, and M. J. Ramírez-Quintana, “El Método Mimético, una Alternativa para la Comprensibilidad de Modelos de ‘Caja Negra,’”. In “Tendencias de la Minería de Datos en España”, ISBN: 84-688-8442-1, pp: 391-402, 2004. [Online]. Available: http://www.lsi.us.es/redmidas/Capitulos/LMD34.pdf.

[19] V. Estruch, C. Ferri, J. Hernández-Orallo, and M. J. Ramírez-Quintana, “Simple Mimetic Classifiers,” in Machine Learning and Data Mining in Pattern Recognition, vol. 2734, P. Perner and A. Rosenfeld, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 156–171. doi: 10.1007/3-540-45065-3_14.

[20] C. J. Mahoney, J. Zhang, N. Huber-Fliflet, P. Gronvall, and H. Zhao, “A Framework for Explainable Text Classification in Legal Document Review,” arXiv:1912.09501 [cs], Dec. 2019, Accessed: Aug. 02, 2020. [Online]. Available: http://arxiv.org/abs/1912.09501.

[21] T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady, “explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning,” IEEE Transactions on Visualization and Computer Graphics, vol. 26, pp. 1064-1074, Aug. 2019, doi: 10.1109/TVCG.2019.2934629.

[22] F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation Forest,” in 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, Dec. 2008, pp. 413–422. doi: 10.1109/ICDM.2008.17.

[23] S. Wang, W. Zhou, and C. Jiang, “A survey of word embeddings based on deep learning,” Computing, vol. 102, no. 3, pp. 717–740, Mar. 2020, doi: 10.1007/s00607-019-00768-7.

[24] C. Raffel et al., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” arXiv:1910.10683 [cs, stat], Jul. 2020, Accessed: Jul. 10, 2021. [Online]. Available: http://arxiv.org/abs/1910.10683.

[25] K. Jaskie and A. Spanias, “Positive And Unlabeled Learning Algorithms And Applications: A Survey,” in 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), Jul. 2019, pp. 1–8. doi: 10.1109/IISA.2019.8900698.

[26] J. Bekker and J. Davis, “Learning From Positive and Unlabeled Data: A Survey,” arXiv:1811.04820 [cs, stat], Nov. 2018, Accessed: Feb. 17, 2020. [Online]. Available: http://arxiv.org/abs/1811.04820.

[27] L. M. de Campos, J. M. Fernández-Luna, J. F. Huete, and L. Redondo-Expósito, “Positive unlabeled learning for building recommender systems in a parliamentary setting,” Information Sciences, vol. 433–434, pp. 221–232, Apr. 2018, doi: 10.1016/j.ins.2017.12.046.

[28] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” Journal of Machine Learning Research, vol. 3, pp. 993–1022, Mar. 2003.

[29] R. Alghamdi and K. Alfalqi, “A Survey of Topic Modeling in Text Mining,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 6, no. 1, 2015, doi: 10.14569/IJACSA.2015.060121.

[30] J. Sahoo, “Modified TF-IDF Term Weighting Strategies for Text Categorization,” Oct. 2018. doi: 10.1109/INDICON.2017.8487593.

[31] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” Machine Learning, vol. 63, no. 1, pp. 3–42, Apr. 2006, doi: 10.1007/s10994-006-6226-1.

[32] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.

[33] W.Cohen, “Fast Efective Rule Induction”. In Proceedings of the Twelfth International Conference on Machine Learning (ICML’95). pp. 115-123. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[34] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, “Image-based Recommendations on Styles and Substitutes,” arXiv:1506.04757 [cs], Jun. 2015, Accessed: Aug. 12, 2020. [Online]. Available: http://arxiv.org/abs/1506.04757.

[35] C. E. Blacke, E. Keogh, and C. J. Merz, “UCI repository of machine learning databases.” University of California, School of Information and Computer Science., 1995.

[36] K. Lang, “Newsweeder: Learning to filter netnews,” in Proceedings of

[9] Funded by Ministry of Science, Innovation and Universities and the European Development Fund (ERDF), in the 2018 call for proposals within the framework of the State Plan for R+D+I Oriented to the Challenges of Society

the Twelfth International Conference on Machine Learning, 1995, pp. 331–339.

[37] J. Mockus, V. Tiesis, and A. Zilinskas, "The Application of Bayesian Methods for Seeking theExtremum," in Toward Global Optimization, vol. 2, Elsevier, 1978, pp. 117–128.

[38] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Computing Surveys, vol. 41, no. 3, pp. 1–58, Jul. 2009, doi: 10.1145/1541880.1541882.

[39] W. L. Taylor, "'Cloze procedure': A new tool for measuring readability," Journalism quarterly, vol. 30, no. 4, pp. 415–433, 1953.

[40] F. Doshi-Velez and B. Kim, "A Roadmap for a Rigorous Science of Interpretability," ArXiv, vol. abs/1702.08608, 2017.

Raúl A. del Águila Escobar

Raúl A. del Águila Escobar, is a PhD. student at the Artificial Intelligence Department of the Escuela Técnica Superior de Ingenieros Informáticos, Universidad Politécnica de Madrid (UPM). He also works as a Data Scientist Specialist at Grupo MasMovil, where he develops and manages Machine Learning projects. His previous working experience was related to Fraud Analytics, having worked as a Manager in EY. He graduated in Computer Science in 2007, and he received an Outstanding Award granted by CEU San Pablo University. He also has a Msc. in Research in Artificial Intelligence by the Spanish Association for Artificial Intelligence and Universidad Internacional Menéndez Pelayo (UIMP). His research areas included knowledge engineering, ontology development and natural language processing. He has presented his PhD. research at the Doctoral Consortium of ECAI 2020.

Mari Carmen Suárez-Figueroa

Mari Carmen Suárez-Figueroa, PhD. is a lecturer at the Artificial Intelligence Department of the Escuela Técnica Superior de Ingenieros Informáticos, Universidad Politécnica de Madrid (UPM) since 2008 and a senior researcher at the Ontology Engineering Group (OEG) since 2002. In addition, she is the Academic Secretary of the Artificial Intelligence Department at UPM since 2017. She graduated in Computer Science in 2001 and got the PhD in Artificial Intelligence, with European mention, in 2010. She has received an Outstanding Award granted by the UPM PhD Commission. Her earlier research lines included ontology development methodologies, ontology evaluation, ontology reuse and ontology design patterns. In these areas, she has participated in several European and Spanish projects (SlideWiki, BuscaMedia, mIO!, NeOn, SEEMP, REIMDOC, OntoGrid, Knowledge Web, PIKON, Esperonto and OntoWeb). Currently, her research is focused (a) on applying artificial intelligence techniques to achieve the so-called cognitive accessibility and (b) on investigating different aspects of the inclusive artificial intelligence. In these research areas she has leaded one internal project at UPM and one bi-lateral project with University of Oxford, and she is the leader of a project funded by Fundación ONCE (ONCE Foundation for Cooperation and Social Inclusion of People with Disabilities). She has been a research visitor at University of Liverpool in 2004, at KMi (Open University) in 2007, at IRIT (Toulouse) in 2012, and at NUIG (Galway) in 2019. She is co-editor of the book "Ontology Engineering in a Networked World" (Springer 2012). In addition, she published her PhD thesis in IOS Press in 2012. She co-organized sessions, conferences, workshops, and tutorials in international events such as ISWC 2019, ISWC 2018, ICTERI 2018, EKAW 2016, ESWC 2014, TKE 2012, ISWC 2012, EKAW 2012, EKAW 2008, ESWC 2008, and WWW 2006.

Mariano Fernández López

Mariano Fernández-López, Phd, is University Lecturer (Profesor Titular) (2013) and the Director of the Degree in Information Systems Engineering at Universidad San Pablo CEU (2014), where he also was the Director of the Software Engineering and Knowledge Engineering Department (2004-2008). Previously, he was Lecturer (profesor asociado) at Universidad Pontificia de Salamanca (UPSAM) (1998-2002) and Universidad Politécnica de Madrid (UPM) (2000-2003). He graduated in Computer Science (1996), obtained a master degree in Software Engineering (1999), a master degree in Knowledge Engineering (2000), and his PhD in Computer Science at UPM (2001). His thesis was awarded the PhD Extraordinary Prize at the Computer Science School. His research areas include ontology engineering, applied formal ontology and knowledge engineering. He is co-author of the book "Ontological Engineering", which received an Honorary Mention for Best Textbook by the UPM University Foundation, and almost 3500 citations according to Google Scholar. He is co-author of 13 book chapters, some of them published by Springer or McGraw-Hill; 7 papers in conferences with proceedings published by ACM, Springer, IOS-Press and IEEE Press; 29 papers in other conferences, workshops and symposia, for instance the Spring Symposium Series of the AAAI-97 (Association for the Advancement of Artificial Intelligence). His paper in IEEE-Intelligence Systems was among the 15 most referenced in the history of that magazine, and, one of his papers in Data & Knowledge Engineering was the most downloaded during 2003. As a member of the research team, he has participated in 21 national and international competitive research projects and two teaching Innovation projects. He has undertaken research stays at University of Sunderland (2001 and 2003), National Research Center of Italy (2001), University of Liverpool (2002) and Open University (2014).

An Efficient Fake News Detection System Using Contextualized Embeddings and Recurrent Neural Network

Junaid Ali Reshi[1]*, Rashid Ali[1,2]

[1] Department of Computer Engineering, Aligarh Muslim University, Aligarh (India)
[2] Interdisciplinary Centre for Artificial Intelligence, Aligarh Muslim University, Aligarh (India)

* Corresponding author. jreshi14@gmail.com

Received 24 February 2022 | Accepted 12 January 2023 | Early Access 9 February 2023

LA UNIVERSIDAD EN INTERNET

Abstract

Fake news is detrimental for society and individuals. Since the information dissipation through online media is too quick, an efficient system is needed to detect and counter the propagation of fake news on social media. Many studies have been performed in last few years to detect fake news on social media. This study focuses on efficient detection of fake news on social media, through a Natural Language Processing based approach, using deep learning. For the detection of fake news, textual data have been analyzed in unidirectional way using sequential neural networks, or in bi-directional way using transformer architectures like Bidirectional Encoder Representations from Transformers (BERT). This paper proposes Contextualized Fake News Detection System (ConFaDe) - a deep learning based fake news detection system that utilizes contextual embeddings generated from a transformer-based model. The model uses Masked Language Modelling and Replaced Token Detection in its pre-training to capture contextual and semantic information in the text. The proposed system outperforms the previously set benchmarks for fake news detection; including state-of-the-art approaches on a real-world fake news dataset, when evaluated using a set of standard performance metrics with an accuracy of 99.9 % and F1 macro of 99.9%. In contrast to the existing state-of-the-art model, the proposed system uses 90 percent less network parameters and is 75 percent lesser in size. Consequently, ConFaDe requires fewer hardware resources and less training time, and yet outperforms the existing fake news detection techniques, a step forward in the direction of Green Artificial Intelligence.

Keywords

Contextualized Embeddings, Deep Learning, Fake News Detection, Natural Language Processing.

DOI: 10.9781/ijimai.2023.02.007

I. Introduction

Fake news has been a buzzword, often heard in journalistic discussions and political discourse. People from various backgrounds use it in different contexts and meanings, as per their understanding, to refer to misinformation, disinformation, rumors, and fake news, etc. There are varying definitions of fake news. Some definitions of fake news are so ambiguous that they eliminate the boundaries between the concepts of fake news, misinformation, disinformation, satire, or even improper and personally offensive news [1], [2]. Among other definitions in the literature, the most consistent definition of fake news is ‘news that is intentionally and verifiably false’ [3]–[5].

Fake news is a phenomenon that can cause serious consequences. These consequences may result in personal, national or global harm. Fake news is shown to spread more quickly than real news [6], and its impact has been studied in various situations, particularly elections [7], [8]. The rapid dissemination of fake news can have serious repercussions. The spread of fake news can cause the democratic processes to be undermined and create chaos. It can create distrust in neutral agencies and spread pseudo-science, thus hurting the communities on a large scale. It has been observed that anti-social elements spread fake news through social media to create law and order problems [8]. Early detection of fake news, is very important in this scenario as it somewhat helps mitigate the ill effects of fake news. It has also been observed that fake news, if encountered by debunking and presenting true news, the retractions fail to completely eradicate the influence of the misinformation[9],[10]. Therefore, a need is felt to detect Fake news at an early stage so that it will not propagate, and as such, we can reduce the harms of fake news spread to a great extent [11].

The problem of mitigating and detecting fake news has been studied by various researchers through different approaches. While some researchers are much concerned about the technicalities in

Please cite this article as:
J. A. Reshi, R. Ali, “An Efficient Fake News Detection System Using Contextualized Embeddings and Recurrent Neural Network”, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 6, pp. 38-50, 2024, http://dx.doi.org/10.9781/ijimai.2023.02.007

the detection and mitigation of fake news on social media, social scientists have been focusing on various psychological aspects of fake news spread and the damage it causes [4], [12]. In empirical studies, many approaches of fake news detection have been experimented with and discussed. Different features have been utilized for the detection of fake news in online social networks through machine learning and deep learning frameworks. As an intelligent system through machine learning needs feature engineering [13], different works have focused on different features. Some of the works have focused on social context features [14], while some have used content-based features [4], [12], [15], [16] to detect fake news. Among other approaches, the propagation aspect of fake news has also been studied and experimented with [17]. Fake news detection using diffusion networks, virality prediction based on network structure, and finding an influential node to determine the dynamics of fake news have also been studied [18], [19]. Fake news detection as a natural language processing problem has also been studied extensively [15], [20].

A. Our Contribution

- This paper proposes a novel fake news detection system – ConFaDe, that uses contextualized word embeddings generated through ELECTRA-based transformer model as an input to LSTM based deep neural network. This model is pre-trained through replaced token detection and masked language modeling tasks.
- The proposed system is evaluated on a well-known benchmark real-world fake news dataset based on the 2016 U.S. presidential elections. It outperforms the existing state-of-the-art (SOTA) Fake news detection system-FakeBERT [36].
- The proposed system uses 90 percent less parameters than FakeBERT and is 75 percent lesser in size than the same. While the FakeBERT model has 135.5M parameters, ConFaDe has only 14.09M. Moreover, ConFaDe utilizes ELECTRA-small as text encoder, which is about ¼ of the size of BERTbase, utilized by FakeBERT as text encoder.
- The proposed system achieves an accuracy of 99.9 percent and an F1 macro of 99.9 percent while training on fewer parameters, consuming fewer resources, and utilizing less hardware, making it an efficient and accurate model for detecting fake news. As the given model intuitively consumes less power, it leads to lesser carbon emission, therefore, taking a step forward in achieving sustainable models for Green Artificial Intelligence (AI).

The rest of this paper is structured as follows: Section II presents an essential background of relevant concepts. Section III contains a review of the pertinent literature. Section IV presents details about the approach and the proposed system. Section V details the experimental setup including the hardware, the software, and various parameters and configurations of the system. Section VI contains the results obtained through experiments and the discussion thereof. The discussion comprises a comparison with baselines, existing works, and the state-of-the-art system on various parameters. Section VII concludes the discussion and provides some future directions.

II. Background

In this section, we discuss various concepts we have used in our set of experiments. We also discuss different approaches to fake news detection and ensuing directions.

A. Word Vector Encodings and Embeddings

To process text, we need to do ample and tailored pre-processing. We cannot have raw text as an input to existing deep learning classifiers. Therefore, we pre-process text and convert it to a vector representation for further processing by deep learning classifiers. Apart from initial pre-processing like stop word removal, stemming, and lemmatization, we need to use a word encoding/embedding technique to create a text vector. There are different types of encodings and word embedding techniques. We briefly discuss GloVe-based embeddings as we further use them in our experiments.

1. GloVe

GloVe, expanded as 'Global Vectors for word representation,' is an unsupervised model for learning vector word representation through training on an aggregated global word-to-word co-occurrence matrix from a large text corpus. It can be used to find word relations like synonyms, antonyms, and other semantic relations like city-capitals, currency-capitals, role-salutation, etc. However, it is not efficient enough to determine word relations such as homonyms [21].

Although there are multiple versions of pre-trained GloVe word embeddings available online, we have used a 300-dimension vector GloVe model trained on 6 billion tokens and 400 thousand words vocabulary from Wikipedia 2014 Gigaword 5 corpus.

2. ELECTRA

'Efficiently Learning an Encoder that Classifies Token Replacements Accurately,' condensed as ELECTRA, is a Bidirectional Encoder Representations from Transformers (BERT) like pre-trained model that generates dense vector representations for natural language tasks. BERT is a Google-developed deep learning framework based on attention mechanism. It is pre-trained on Masked Language Model (MLM) and Next Sentence Prediction (NSP) tasks, useful for various downstream natural language processing tasks [22]. There are many versions of BERT, depending on the sentence length, corpus trained on, number of encoders, and number of attention heads used. In its pre-training, ELECTRA partially replaces the MLM task in BERT with the replaced token detection task (RTD) [23]. Somewhat like Generative Adversarial Networks (GANs), albeit with maximum likelihood, and not adversarial training, the ELECTRA model is trained to discriminate between 'real' and 'fake' input data, which is infused by corrupting some input tokens with plausible 'fake' tokens. The pre-training task requires the discriminator part of the model to determine the corrupted or intact tokens, which are fed by a small generator network. The addition of the RTD task has led to improvement in the model's performance for a given size, computing power, and data [23].

III. Literature Review

Machine Learning and deep learning find their application in most modern-day intelligent applications. Similarly, machine learning and deep learning models have found extensive usage in the task of fake news detection [24] [25]. With Machine Learning, feature engineering is the most crucial step, where working on the features is an essential step in improving the performance of a model [26], [27]. In early studies, to solve the problem of fake news detection through machine learning, a manual set of features were designed along with lexical and syntactical features [28], [29]. Among other machine learning algorithms, Random forest (R.F.), Support Vector Machines (SVM), and Naïve Bayes (N.B.) classifiers have been extensively used to detect fake news [1], [20], [24], [25], [30]. In machine learning models, apart from the pre-processing, text requires a lot of feature engineering before it is ready as an input. With the advent of deep learning driven technologies, deep learning models gradually replaced machine learning approaches. It was predominantly due to the fact that unlike machine learning models, deep learning models do not require explicit feature engineering to perform well. The input to these deep learning frameworks can be in the form of text, images, or videos, depending on the type of fake news detection approach [31]. In multimodal fake news detection, multiple architectures are combined to form a hybrid architecture, which is then used to detect fake news in text, images, and videos [32]. For text, the content is embedded at the word or sentence level [33]; for image, a pixel-based tensor is used as an input form to the deep learning framework [32]. After the preparation of input tensors, many neural network based architectures like Convolution Neural Networks (CNNs) [34]–[36] and Recurrent Neural Networks (RNNs) have been used. CNNs are primarily used to extract features from the text in a more efficient way. CNNs with average or max pooling find their use in fake news detection tasks. Recurrent Neural Networks such as Gated Recurrent Units (GRUs) [37], LSTMs [38], [39], and Bidirectional recurrent neural networks (BRNNs) find great utility in text processing [40]. GRUs contain only two gates –reset and update and as such are easier to use than LSTM, which consists of an input gate, an output gate, a forget gate, and a cell. While LSTM can recall short-term memories for a long time, so as to aid in forming the context, it still processes the text in a unidirectional way[38]. It has also been observed that preference of GRUs over LSTM or LSTM over GRUs is decided by the computational resources available[41]. Recently, many works have focused on using transformers for extracting multimodal features of news content [42], [43].

In terms of the type of features used for fake news detection, many researchers approach the problem of fake news detection as different tasks like News content detection problem [4], [6], [44], News diffusion dynamics [45]–[47], and/or social context based problem[14]. As Social media content is primarily composed of text, news content-based features focusing on text are much helpful in fake news detection [48], [24].

In the existing literature, various configurations of deep learning networks combined with different features have been used to mitigate the fake news detection problem. These deep learning methods have been preferred as they are able to capture different patterns in text (news) implicitly, which machine learning algorithms are not capable of capturing until explicitly engineered. As raw text still cannot be placed as an input to the deep learning architecture, different word encoding and embedding schemes have been used as the first step[49]. Additionally, many efforts have been made to make programming configurations easier to implement and reduce the computational power needed [50]. In one of the pioneering works of using a deep learning framework for detecting fake news, Ma et al. use basic tf-idf based text encoding in their deep learning architecture for detecting fake news [51]. Their model performed better than many state-of-the-art (SOTA) machine learning based models at the time [51]. Many text embeddings have been used to simplify the problem of fake news detection. Document-level embeddings [52], Sentence-level embeddings [52], [53], and word-level embeddings have been utilized by various researchers to generate input vectors for a Machine learning or deep learning classifier [1]. Some studies use word-level embeddings like Word2Vec [54] and GloVe [55] to obtain text vectors for subsequent use by deep neural networks. Many language models that are based on RNNs and Transformers, like Embeddings from Language Models (ELMo) [56], FastText [57], and BERT [36] [58] have also been utilized for the generation of text embeddings. These text embeddings have been further used in deep learning models for fake news detection [58]–[61]. However, the problem with huge models like BERT is that it requires a lot of computational power and time to use them for downstream tasks, and thus, in the long run, we look for a better alternative.

Further, among deep learning models, some researchers have used CNNs, some have used RNNs, while some have used ensemble approaches for developing the frameworks to detect fake news [12], [16], [45]. Irrefutably, there has been a quest for getting the right approach in selecting a text embedding along with a neural network, as both are essential to the performance of a model/system. Improvement in either of these two research areas can lead to better systems for fake news detection. The recent techniques in fake news detection include use of Graph Neural Networks (GNN), Generational Adversarial Networks (GANs) and ensemble approaches. Graph neural networks operate on graph structure by recursive node classification. They process the global structural features better than other deep learning architectures [62]. Graph Convolutional Networks (GCN) and Propagation Graph Neural networks (PGN) are some important techniques that belong to Graph Neural Networks [63]. Generative adversarial networks are also used for fake news detection, albeit for images and videos only [63]. Adversarial training is done to generate synthetic fake images and videos. DeepFake is an area of applications of GANs where fake news detection can be indirectly achieved with the help of GANs [64]. Ensemble methods are created by combining several models for performing a single task at the end. CNN+LSTM ensembles have been used frequently for fake news detection [63]. Ensembles based on Bi-LSTM + CNN, RNN + SVM, Attention mechanism+RNN and other configurations have been tried in the literature [65], [66].

IV. Methodology

In this section, we discuss the proposed fake news detection system, starting from data preparation and culminating by describing the model in detail.

The core objective of this study is to develop an efficient and accurate fake news detection system. By using fewer resources, the system should be capable of detecting fake news with an accuracy that it may outperform the state-of-the-art system on this task. For its accuracy, we evaluate the system using different performance measures, as explained in the Section V.D. For measuring the efficiency, we compare the system with an already existing state-of the-art (SOTA) system in terms of parameters and resources used (details in Section VI.C).

We propose an LSTM-based model that leverages contextualized embeddings generated from a transformer-based architecture-ELECTRA, to detect Fake News. We call this model as ConFaDe. The process of fake news detection is illustrated in Fig. 1.

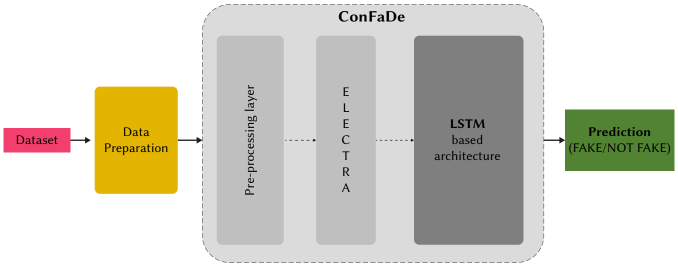


Fig. 1. Process of Fake News detection in the proposed system.

As seen in Fig. 1, we prepare the dataset at the first step of the process. The prepared dataset is then fed to ConFaDe architecture, which classifies the news as fake or not fake. The detailed process of data preparation is discussed in Section IV.A. Inside ConFaDe, before providing input to the LSTM based architecture, we pre-process the data through a layer to make data ready for our ELECTRA architecture. Subsequently, an LSTM based deep learning architecture classifies the news as ‘fake’ or ‘not fake’. The detailed architecture of the proposed model is discussed in Section IV.C. Our model uses the word embeddings generated through the ELECTRA-Small++ model, pre-trained on a large uncased English text corpus by Google Research.

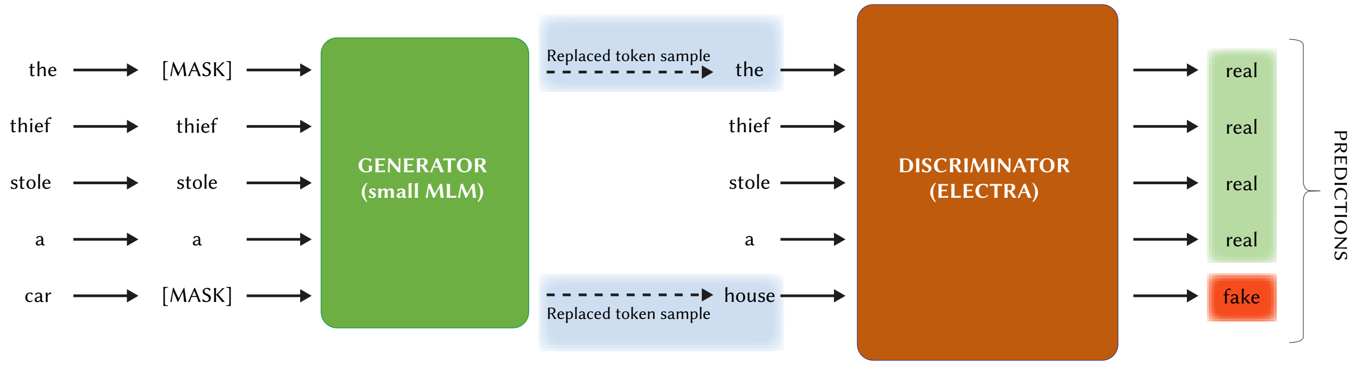


Fig. 2. Exemplifying Replaced Token Detection and prediction in ELECTRA.

A. Data Preparation

Benchmark representative datasets are a standard in evaluating performance of a system [67],[68], [69], [70]. We select a well-known benchmark dataset of fake news, which contains fake and real news propagated during 2016 U.S. Presidential elections [36], [68], [69]. The raw input file is first examined for inconsistencies. Values that are in Arabic or are not legible and do not contain English language are removed, as our transformer model is trained on English corpus. After removing such values, we replace 'Null' and 'nan' values with blanks. We further perform initial pre-processing on text, removing punctuations, stopwords and URLs. After this process, we are left with 20718 labeled instances. We further split the dataset in the ratio of 90:10 for Training + validation (18646 instances) and testing (2072 instances). Of the 18646 instances, 13053 were used for training, and 5594 were used for validation, in the ratio of 70:30.

B. ELECTRA Training and Hyperparameters

The ELECTRA model is trained using two encoder-based neural networks, Generator (G) and Discriminator (D). The first one is a Generator G, which maps a sequence of input tokens $\boldsymbol{a} = [a_1, a_2, \dots , a_n]$ into contextualized vector representations $c(x) = [c_1, c_2, \dots, c_n]$. The Generator is trained to perform masked language Modeling (MLM). MLM randomly selects random positions (with integer values running from 1 to n) to mask out the original input $m = [m_1, m_2, \dots, m_k]$. When the positions are fixed, the corresponding tokens are replaced by a [MASK] token:

$$a^{masked} = REPLACE(a; m; [MASK]) \tag{1}$$

The masked out tokens are then replaced by generator samples:

$$a^{corrupt} = \mathrm{REPLACE}(\mathrm{a}; \mathrm{m}; \hat{\mathrm{a}}) \tag{2}$$

Where $\hat{a}$ represents a plausible generator sample and is given by

$$\hat{a}_i \sim p_G(a_i|a^{masked})\ for\ i \in m \tag{3}$$

and model inputs $\mathrm{m_i}$ are constructed as:

$$m_i \sim unif\{1, n\}\ for\ i = 1\ to\ k \tag{4}$$

The probability of generating a token a_t with a softmax layer at the Generator is given by:

$$\mathrm{P_G(a_t|a) = exp(e(a_t)^T c_G(a)_t) / \textstyle\sum_{a'} exp(e(a')^T c_G(a)_t)} \tag{5}$$

where t is a given position, e the token embeddings, and all other expressions hold the usual meaning. The Generator is specifically trained to increase the likelihood of masked out tokens and is not supplied with noise, like in adversarial training.

For a given position t, the discriminator D predicts whether the token a_t is corrupted or not, i.e., whether it is from the Generator and not the original data distribution, in specific terms; whether $\mathrm{a^{corrupt}}$ matches the original input a.

$$\mathrm{D(a, t) = sigmoid(w^T c_D(a)_t)} \tag{6}$$

The whole process of Masked Language Modeling, Replaced Token Detection, and prediction is illustrated through an example in Fig. 2.

In Fig. 2, the text 'the thief stole a car' is the input text in which some tokens are replaced with [MASK] token. The Masked-out tokens are replaced (corrupted) by Generator samples and provided as input to the Discriminator. The Discriminator then predicts whether the given token is 'fake' (corrupted/replaced) or 'real' (not corrupted/original). The token 'house' is a corrupted token, the original being 'car'. The ELECTRA model in the example predicts the last token 'house' as 'fake' (corrupted) and the rest as 'real'.

After pre-training, Generator is not used, and the Discriminator is trained on the downstream tasks. In our case, it was pre-trained on a large uncased English text corpus by Google Research. The text corpus is not public and hence not available for any experimenting. The combined loss in the model is minimized as:

$$\min_{\theta_G} \theta_D \ \textstyle\sum_{a \in X} \ L_{MLM}(a, \theta_G) + \lambda L_{Disc}(a, \theta_D) \tag{7}$$

Where loss function of MLM is given as:

$$L_{MLM}(a, \theta_G) = \ \mathbb{E}\left(\textstyle\sum_{i \in m} -\log p_G\,(a_i|a^{masked})\right) \tag{8}$$

and loss function of Discriminator is given as:

$$L_{Disc}(a, \theta_D) = \ \mathbb{E}\begin{pmatrix} \sum_{t=1}^{n} 1(a_t{}^{corrupt} = a_t) \log D(a^{corrupt}, t) \\ + \\ 1(a_t{}^{corrupt} \neq a_t) \log 1 - D(a^{corrupt}, t) \end{pmatrix} \tag{9}$$

Where the symbols used hold their usual contextual meaning.

The pre-trained model used in our work has an additional dense layer on the top of CLS token and has been initialized by an identity matrix. The fine-tuned hyperparameters for the training of ELECTRA Small are given in Table I.

TABLE I. Fine-tuned Parameters for ELECTRA Small

Hyperparameter	Value
Batch size	32
Learning Rate	1e-4
Adam β$_1$	0.900
Adam ε	1e-6
Adam β$_2$	0.999
Learning rate decay	Linear
Layer-wise L.R. decay	0.800
Dropout	0.100
Attention dropout	0.100
Warmup fraction	0.100
Weight Decay	0

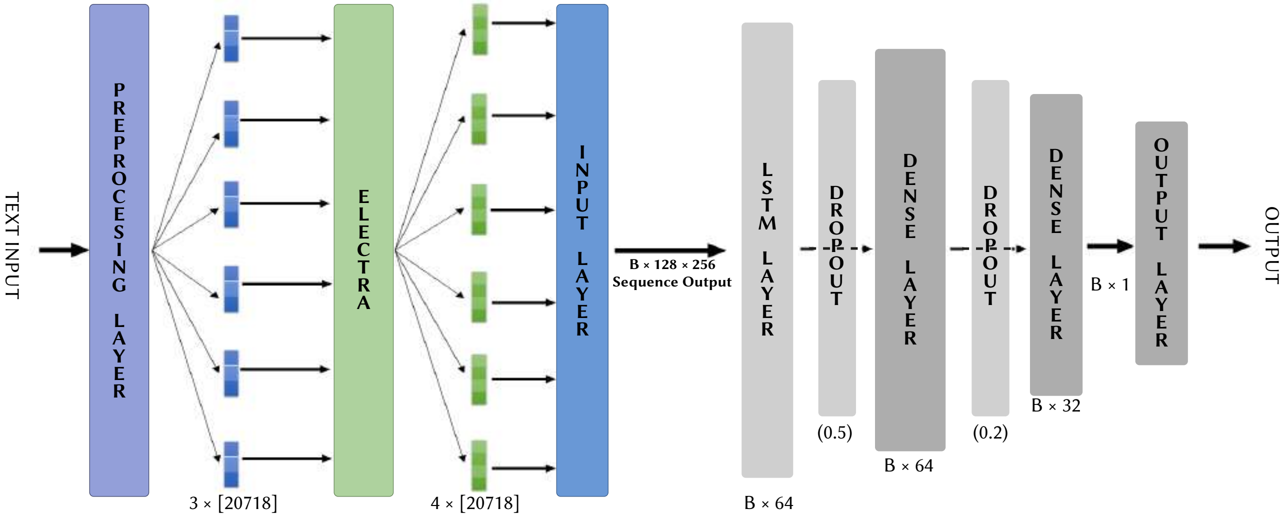


Fig. 3. Layered Architecture of ConFaDe.

TABLE II. ConFaDe Layered Architecture

Layer	Input Dimensions	Output dimensions	Parameter number
Pre-processing	Plain Text 20718 rows	Dict[3]×Tensor[20718]	-
ELECTRA	Dict[3]×Tensor[20718]	Dict[4]×Tensor[20718]	110M
Input	Dict[4]×Tensor[20718]	128×256	-
LSTM	64×128×256	64×64	82176
Dropout	64×32	64×32	0
Dense	64×64	64×64	4160
Dropout	64×32	64×32	0
Dense	64×64	64×32	2080
Dense	64×32	64×1	33

C. Proposed Model: ConFaDe - The Fake News Detection System

The proposed model - ConFaDe consists of a transformer-based model for generation of text embeddings and LSTM based deep learning architecture for further classification of news as fake or not. The details of the ConFaDe architecture are shown in Table II and illustrated in Fig. 3.

Table II lists different layers with input dimensions, their output dimensions, and the number of parameters. It can be observed that the parameters in different layers of classifier architecture are limited. The batch size in the following experiments is set to 64. For the sake of explanation and visualization, the layered architecture of ConFaDe is illustrated in Fig. 3.

As can be seen from Fig. 3, ConFaDe consists of various layers, the description of which is as follows:

Pre-processing layer: This layer is used to pre-process text. It maps a string Tensor to a dictionary of numeric tensors, which is the required input for the transformer architecture. It performs the basic operations on the text, including some pre-processing. It provides, as an output, a dictionary of numeric values mapped to the text string of the shape of batch size, as defined by the architecture.

ELECTRA layer: This layer accepts a dictionary of tensors as an input and performs training on the tensors for embedding computation. It returns a dictionary of computed outputs on the text. There are four outputs in this layer. The output under key 'Pooled_output' contains the embedding for each sentence as it appears in the corpus and is two-dimensional. The 'Sequence_output' provides contextualized word-level embedding for each sentence limited to the maximum word length. It is three-dimensional. The output under 'encoder_output' provides the output from each encoder. It is noticeable that the last layer of encoder output is actually the sequence output as it logically should be. Under the key 'default' is a Tensor of shape [batch_size× dimension].

Input layer: This layer is used to prepare input for the subsequent deep learning layer architecture. It gives, as an output, sequence_output of the previous layer to be processed by the LSTM layer. It performs no other operations on the data.

LSTM layer: LSTM is a kind of recurrent neural network, which has the capability of learning long-term dependencies [39]. The hyperparameters used in the architecture of LSTM are listed in Table IV. With 64 as batch size, it takes 128×256 shaped tensor and provides a tensor with shape as 64 to the next dense layer.

Dense Layer: A layer of neurons connected together, with each neuron receiving input from the previous layer, is called a dense layer. A dense layer is capable of learning representations based on the input. A dense layer, in essence, carries out matrix-vector multiplication and provides an output as the application of activation function to dot product of input (data) and kernel (weight matrix), with the addition of bias. The activation function used in these layers is Rectified Linear Unit (ReLU).

Dropout: The dropout layer is used for the regularization of a neural network, to avoid overfitting. During training, some random neurons are ignored by not activating them in the forward pass and not updating their weights in the backward pass so that every neuron contributes to learning, and only some neurons may not remember the pattern. In ELECTRA pre-training, the dropout is set to 0.1. In dense layers, it is set to 0.5 and 0.2.

Output layer: This layer consists of the neurons, which are fired according to the prediction on the data. This layer consists of a different activation function, 'Sigmoid', as the predictions are not continuous but binary.

Activation functions: This function is a non-linear transformation applied to the inputs from the previous layers to provide output. In our model, we use Sigmoid activation function at the output layer and ReLU as an activation function in other layers.

Loss function: It is a function to calculate the gradients for updating weights in a neural network. We have used binary cross-entropy loss function, which is mathematically calculated as:

$$Loss = -\frac{1}{output\ size}\sum_{i=1}^{output\ size} \alpha_i \ .\log\widehat{\alpha_\iota} + (1-\alpha)\ .\log(1-\ \widehat{\alpha_\iota}) \qquad (10)$$

Where α_i is the target value, $\widehat{\alpha_\iota}$ is the value of i^{th} scalar, and the output size is the number of scalar values in the model output.

Optimizer: This function is used to update model parameters like weights and learning rate, and minimize loss functions to achieve maximum performance in a deep learning algorithm. We have used Adam optimization algorithm, a version of the stochastic gradient descent method based on adaptive estimation of lower order (first and second) moments [71].The parameters used are ε = 1e-08, decay=0.0, beta_1=0.9, learning rate=0.001, and beta_2=0.999.

V. Experimental Setup

We carried our experimentation using DELL PowerEdgeR740 Server P.C. with Intel Xeon Silver 4114 CPU with 20 core(s). We use NVIDIA Quadro P4000 GPU with 1972 CUDA cores, peak single precision of 5.3 TFLOPS, DDR5 memory of 8 G.B., Memory bandwith of 243 Gb/S, Memory Interface of 256 Bits, and a Maximum power consumption of 105W. We trained the model using Tensorflow, and Python 3.8.8, and CUDA version 11.4.

A. Dataset Description

The dataset that has been used for experimentation is openly available and has been used by various researchers in their experiments [36]. It contains a collection of labeled fake news and real news propagated during the U.S. General presidential Election - 2016. The dataset can be downloaded from the internet. It comprises two data files:

(i) train.csv: This file contains training data with the following attributes:

- id: unique id for a news article
- title: the title of a news article
- author: author of the news article
- text: the text of the article
- label: label of the corresponding news article, having two values as:
    - 1: Fake.
    - 0: Real.

(ii) test.csv: This contains testing data with no labels.

After initial data preparation, as earlier explained, the dataset contains 20718 instances, the description of which is listed in Table III.

TABLE III. Dataset Description

Feature	Number of Instances (raw dataset)	Number of instances (Processed dataset)
id	20800	20718
title	20242(excluding missing, including null)	20160
author	18843(excluding nan)	20718
text	20761(excluding missing)	20679
label	20800	20718

The pre-processed dataset consists of 10369 instances with class '1' and 10349 instances of class '0'. The ground truth has been labeled by the contributor of the dataset. The detailed process of collection of dataset is not available. As this dataset has been used extensively in the literature, we also use it in our experiments.

B. Experimental Configurations

We conducted various experiments with different embedding and architectures to present a baseline. We categorize the experiments based on the type of embeddings used. We use different types of encodings and embeddings. We start with one hot encoding, then move on to integer encoding. After that we try GloVe pre-trained embeddings, and at last we use ELECTRA generated encodings. We classify our experiments as transformer based models and non-transformer based models, depending upon the use of transformers in any sub-task. We try different configurations of these models and report the performance of only those models whose performance was best and had comparatively few parameters. The model hyperparameters for these models were set to default and the batch size was set to 64 for all of them. The Training: Validation: Testing split was the same for all of the experiments (as explained in Section IV.A). The evaluation metrics used are detailed in sub-section V.D.

C. Model Hyperparameters

Hyperparameter selection is one of the most important aspects of a deep learning architecture. Optimal hyperparameters are very important for a deep learning framework to perform well, while reducing cost and memory utilization. For manually selecting optimal hyperparameters, knowledge of the problem, domain, and deep learning is required. Table IV provides the optimal hyperparameters used in our LSTM architecture.

TABLE IV. Hyperparameters of LSTM Based Deep Learning Architecture (ConFaDe)

Hyperparameter	**Values**
Dropout rate	0.5, 0.2
Activation function	ReLU, Sigmoid
Learning rate	0.001
Loss function	Binary Crossentropy
Optimizer	Adam
Batch size	64
No of epochs	13
Recurrent activation	Sigmoid
Recurrent_initializer	orthogonal
Bias	True
Bias_initializer	zeros
Kernel_initializer	Glorot_uniform
Recurrent Dropout	0
Unit_forget_bias	True

D. Evaluation Parameters

To evaluate the performance of a classifier for the task at hand, various performance metrics or evaluation parameters are used. For a classification task, the confusion matrix is an important performance measure. In multi-level classification, it clearly represents the number of classifications or miss-classification an algorithm does by assigning the number of instances that the algorithm thinks to belong to a particular class versus the actual class the instance belongs to, in a tabular format. In binary classification problems, the confusion matrix is an important performance depicter. In the context of fake news detection, the confusion matrix consists of the following values:

True Positive (T.P.): When the algorithm characterizes a news article as fake when it is actually labeled as fake.

False Positive (F.P.): When the algorithm characterizes a news article as fake when it is actually labeled as true.

False Negative (F.N.): When an algorithm characterizes a news article as true when it is actually labeled as fake.

True Negative (T.N.): When the algorithm characterizes a news article as true when it is actually labeled as true.

Apart from these, different metrics are used in evaluating the performance of classifiers [72], which are as below:

- **Accuracy:** It gives a measure of similarity between predicted fake news and actual fake news.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

- **Precision:** It measures the objective of the classifier, which here is to detect fake news, and quantifies the fraction of all predicted fake news that is actually labeled as fake news. A value of precision closer to 1 or 100% is best. This measure is often used in conjunction with the Recall, as the precision will automatically be high with few positive (fake news) predictions. A classification model that does not produce any false positives has the maximum value for precision.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{12}$$

- **Recall:** It quantifies the sensitivity measure or the fraction of positive (fake news) articles that are categorized/predicted as fake news. It gives the measure of the degree of correctness of a classifier with respect to predicting only a particular class (positive/fake) and does not take into consideration the false positives. A model with no False negatives will have a maximum value of 1 or 100% for the Recall.

$$\text{Recall} = \frac{TP}{TP+FN} \tag{13}$$

- **F1 Score:** It is also called as F-measure. It is the harmonic mean of precision and Recall, which provides an overall measure of the prediction performance of a classifier in predicting fake news.

$$F - measure = 2\ \frac{precision * recall}{precision + recall} \tag{14}$$

- **F1-Macro:** It is almost the same as F-measure but differs only in how it is calculated. For a binary or multi-class classification, when we tend to take precision, Recall, and F measure of individual classes, for a total performance measure, we use F1-macro. It is defined as the mean of the class-wise F-measure values and gives equal weightage to all classes (fake and real) in the dataset [73].
- **FNR:** Also called as miss rate. It gives a measure of how many fake news articles were misclassified by the classification algorithm.

$$\text{False Negative Rate (FNR)} = \frac{FN}{TP+FN} \tag{15}$$

- **FPR:** Also called as fall-out, it is the proportion of the negative classes (real news) identified as positive classes (fake news).

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP+TN} \tag{16}$$

VI. Results and Discussion

In this paper, we experiment with various text embeddings and deep learning models for fake news detection. We particularly investigate the use of contextualized embeddings for fake news detection with the help of an LSTM based deep learning model. We use ELECTRA for generating text embeddings and name this model as ConFaDe. The ConFaDe model performs better than the existing best models with an accuracy of 99.9 % and consumes less time and resources than the existing state-of-the-art models. We also conducted some experiments with different vector representations and embeddings to draw a comparison. We used DNN based classification model and LSTM based classification model with different embeddings. The confusion matrix for each experiment is given in the corresponding section, and performance on different evaluation measures is given in appropriate section.

A. Non-Transformer Based Models

We conducted several experiments with deep learning models and different word embeddings. We use a simple deep neural network (DNN) and an LSTM based network to estimate the overall efficiency in tandem with different word embeddings. For a better comparison, we start with one-hot encodings at the word level. After that, we use integer encodings with an embedding layer for learning word embeddings.

With one hot encoding, vocabulary built on top-512 words, the classifier performs just fine. This gives an intuition that the news headlines may play an important role in helping to classify fake news. However, this conclusion is not definitive, as the limitation of vocabulary leads to the loss of much information and the vectors are too sparse when built through the strategy. With the accuracy of 72.15% and 74.18 % of the basic DNN and LSTM based network respectively, the models did not achieve outstanding results. The Recall in both cases is low, with DNN having a recall of 59.7 % for fake class and LSTM having a Recall of 66.67%. This points out to the fact that the classifiers missed many of the fake classes. The corresponding confusion matrices are listed in Table V and Table VI.

TABLE V. Confusion Matrix for One Hot Encoding With DNN

	Predicted Fake	Predicted True
Actual Fake	643	434
Actual True	143	852

TABLE VI. Confusion Matrix for One Hot Encoding With LSTM

	Predicted Fake	Predicted True
Actual Fake	719	358
Actual True	177	818

With integer encoding, maximum length fixed at 512 words, and vocabulary limited to 5000 words, we build DNN and LSTM models. We use an embedding layer with an output dimension of 100 to learn the encoding. The performance of the DNN based model and the LSTM based model with integer encoding is observable from the confusion matrix in Table VII and Table VIII, respectively.

TABLE VII. Confusion Matrix for Integer Encoding With DNN

	Predicted Fake	Predicted True
Actual Fake	1029	48
Actual True	55	940

TABLE VIII. Confusion Matrix for Integer Encoding With LSTM

	Predicted Fake	Predicted True
Actual Fake	1001	76
Actual True	70	925

The DNN based model shows an accuracy of 95.02%, and the LSTM based model shows an accuracy of 92.95%. The F1 Macro, indicative of the overall performance of DNN based model and LSTM based model, is also impressive, with values equal to 95.01% and 92.94%, respectively. It implies that the encoding approach of creating an embedding layer to learn the embeddings is a better method than one-hot encoding in this case.

We also use a pre-trained word embedding- GloVe, with an embedding dimension of 300, to observe the behavior of the classification algorithms. As explained earlier, it uses 6 billion tokens and has a dimension of 300. With classification models using a static embedding matrix initialized with the GloVe embedding, and the layer set to non-trainable, both DNN and LSTM based models had a comparable accuracy of 95.02%, with their F1 scores being 93.70% and 93.23% respectively. The Confusion matrix for models using Glove embeddings in conjugation with DNN and LSTM are listed in Table IX and Table X, respectively.

TABLE IX. Confusion Matrix for Glove Embedding With DNN

	Predicted Fake	Predicted True
Actual Fake	1025	52
Actual True	78	917

TABLE X. Confusion Matrix for Glove Embedding With LSTM

	Predicted Fake	Predicted True
Actual Fake	1008	69
Actual True	71	924

B. Transformer Based Models

The transformer-based models for generating word representations have shown promise in solving downstream tasks. We use ELECTRA based transformer model to generate word embeddings for the sentences. The word embeddings we generate are contextual in nature and quite powerful for downstream tasks.

a) Simple DNN Models

We train a simple Deep Neural Network (DNN) with two dense layers of size 64 and 32 and an output layer. Fig. 4 contains the loss curve as observed during the training of the DNN based model.

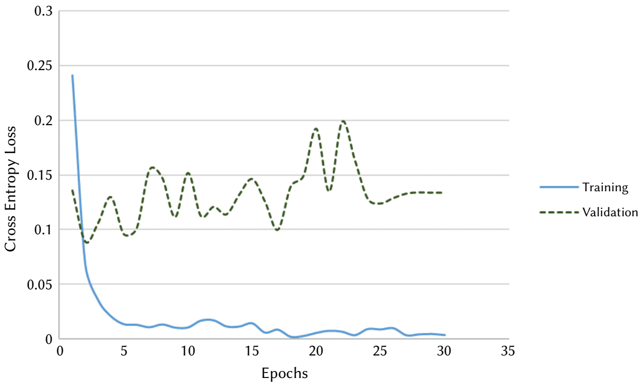


Fig. 4. Loss graph of ELECTRA based DNN based model.

It is pretty clear by observing the loss curve in Fig. 4 that the training loss and validation loss, both are converging, indicative of model learning. As it can be observed that the training loss is almost constant at epoch 30, validation loss also becomes somewhat steady and lowest around epoch 30. We use early stopping criteria for training as no substantial decrease in training loss for five consecutive epochs (patience=5) and minimum delta=0.0001. On the mentioned configuration, the optimized time taken for training the model was about 3 minutes (2.7 minutes) and the whole process takes about 1.15 hours (excluding the time for optimizing the environment). Fig. 5 contains the accuracy plot of ELECTRA based DNN model.

In Fig. 5, we can observe the accuracy curve of training and validation of simple DNN based model. By observing the accuracy plot of the ELECTRA based DNN model, it can be noted that the validation accuracy and the training accuracy tend to converge and achieve a plateau around epoch 30. The graphs are indicative of the point that the model is not overfitting on the data. The Confusion matrix obtained during model testing is given in Table XI.

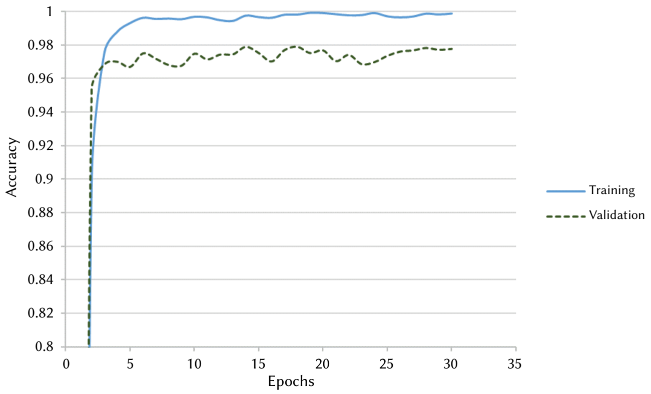


Fig. 5. Accuracy of ELECTRA based DNN based model.

TABLE XI. Confusion Matrix for ELECTRA Based Embedding With DNN

	Predicted Fake	Predicted True
Actual Fake	1055	22
Actual True	25	970

The DNN based model, when used with the ELECTRA-generated word embeddings, had an accuracy of 97.73% and an F1 score of 97.78%.

b) LSTM Based Model (ConFaDe)

On using LSTM based neural network model and trying different configurations, the best configuration yielded an accuracy of 99.9% with an F1 score of 99.9%. On examining the training graph of the ConFaDe, we can see that the validation loss and training loss almost converge near the 13th epoch. This is the optimal point for stopping the model training.The Loss and accuracy curves associated with the model are given in Fig. 6 and Fig. 7, respectively.

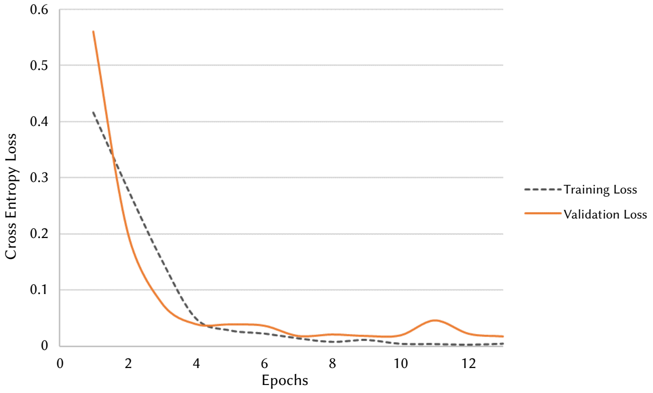


Fig. 6. Loss graph of ConFaDe model.

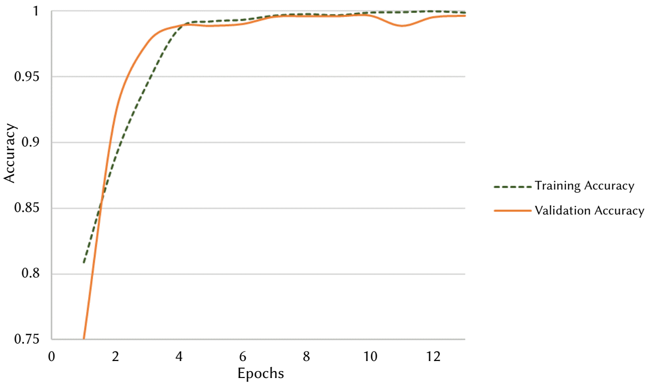


Fig. 7. Accuracy graph of ConFaDe model.

As observable from Fig. 6, the model sees a steady decline in both the training loss and the validation loss, and achieves a plateau around 13 epochs, the stopping criteria being no further substantial decrease in loss for 3 epochs (patience=3, min_delta=0.0001).

From the accuracy plot of training and validation of ConFaDe in Fig. 7, the curve is indicative of model achieving maximum learning at the 13th epoch, as the validation accuracy and training accuracy almost converge. These graphs, when combined together, are also indicative of the fact that the model has not overfit on the data.

For obtaining the actual performance metrics of the model, we run the model on unseen test data, as already explained earlier, to get the results. The confusion matrix obtained by testing the model on the unseen data is given in Table XII.

TABLE XII. Confusion Matrix for ELECTRA Based Embedding With LSTM (ConFaDe)

	Predicted Fake	Predicted True
Actual Fake	1075	2
Actual True	0	995

As is evident from the confusion matrix in Table XII, the total number of False Positives is 0, and that of False Negatives is 2. This is self-explanatory of the performance of the model as it does not create a false alarm and has a very low miss rate.

To evaluate the performance of our model in detail, we have used various performance measures like precision, Recall, and F1 score of individual classes, and F1 macro and Accuracy of different models. Table XIII presents the details of the performance of each model.

It is quite evident from the reported metrics in Table XIII that the model outperforms other model configurations on the same dataset with less loss, more accuracy, Precision, Recall, and F1 score.

To give an idea of how the performances vary with the change in embeddings, we also provide a comparison between the accuracy and F1-macro in different models, using different embeddings in Fig.8 and Fig. 9, respectively.

As we know, LSTM processes the sequence of text singularly to remember context of words. Therefore, it is expected to show some improvement in performance when augmented with a better encoding scheme. DNN just remembers the patterns as a whole and doesn't include any context. A better encoding scheme which captures context in a better way may also increase the performance of DNN. One hot encoded vectors suffer from sparsity problem. Integer encoding performs well because we create our own embedding matrix in it and it is also trained to learn better representations. Nevertheless, it may also require more data to perform better. We also use pre-trained GloVe embeddings for both the models. As we set the embedding static, the performance in both the models seems alike. At the end, we use ELECTRA to generate contextual embeddings which learn context of whole sentences both ways, and at once. The results in the performance of both the models is excellent as compared to other baselines.

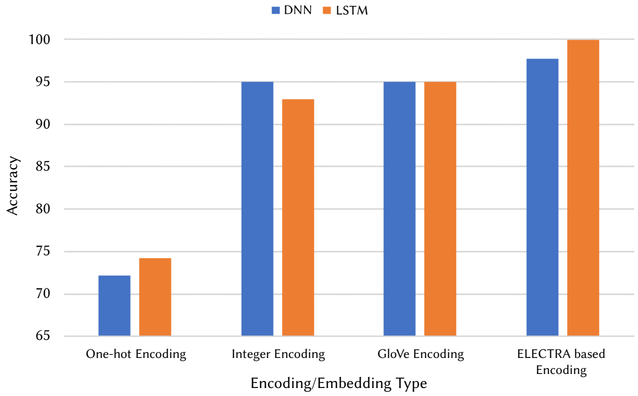


Fig. 8. Comparison of accuracy of different model configurations.

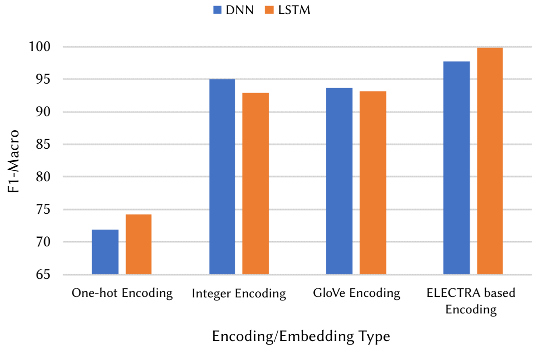


Fig.9. Comparison of F1-Macro of different model configurations.

TABLE XIII. Performance Metrics of Various Classifiers and Vector Representations

Vector representation	Classification Model type	Fake Class(1)			Real Class(0)			Accuracy	F1 Macro
		Precision	Recall	F1	Precision	Recall	F1		
One hot encoding	DNN	81.8	59.7	69.03	66.25	85.63	74.70	72.15	71.87
One hot encoding	LSTM	80.25	66.76	72.88	69.56	82.21	75.36	74.18	74.12
Integer Encoding	DNN	94.92	95.54	95.23	95.14	94.47	94.80	95.02	95.01
Integer Encoding	LSTM	93.64	92.94	93.20	92.40	92.96	92.68	92.95	92.94
GloVe	DNN	92.92	95.72	94.04	94.63	92.16	93.38	95.02	93.70
GloVe	LSTM	93.41	93.35	93.50	93.05	92.86	92.95	95.02	93.23
ELECTRA	DNN	97.69	97.96	97.82	97.78	97.49	97.63	97.73	97.78
ELECTRA	**LSTM**	**100**	**99.81**	**99.90**	**99.80**	**100**	**99.89**	**99.90**	**99.90**

C. Comparison With Previous Works

We compare our model - ConFaDe with different models that exist in literature. We also make comparison of our model with state-of-the-art (SOTA) model- FakeBERT on different parameters, as listed in Table XIV.

From Table XIV, it is evident that ConFaDe architecture uses a smaller number of parameters than the other transformer-based architecture -FakeBERT. ConFaDe architecture has only 14M encoding parameters, in comparison to 110M encoding parameters of FakeBERT architecture. In addition, ConFaDe has only 88.4K parameters in its classification model whereas FakeBERT has 25.5M parameters in its classification model. In entirety, FakeBERT has 135.5M parameters while ConFaDe only has 14.09M parameters.

On comparing our work with the previous BERT based SOTA model, it can be seen that our model performs better. Fig 10 and Fig 11. provides the comparison of cross entropy loss and accuracy of the two models.

TABLE XIV. Comparison of ConFaDe With FakeBERT

PARAMETER NAME	FakeBERT	ConFaDe
Number of Transformer blocks/layers	12	12
Encoding Hidden Size	768	256
Encoding model Parameter number	110M	14M
Classification model Parameter number	25.5M	88.4K
Total Parameter number	135.5M	14.09M

From Fig. 10 and Fig. 11, it can be observed that our model has less cross entropy loss (binary) than FakeBERT based model and its accuracy is higher than SOTA BERT based model – FakeBERT. In Fig. 12 and Fig. 13, we present the comparison with FakeBERT with False Negative Rate (FNR) and False Positive Rate (FPR) as performance measures.

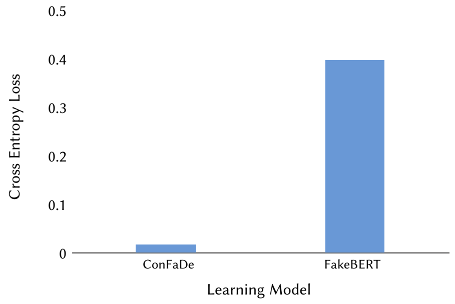


Fig. 10. Loss comparison with FakeBERT.

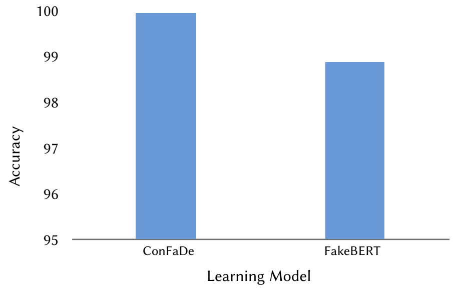


Fig. 11. Accuracy comparison with FakeBERT.

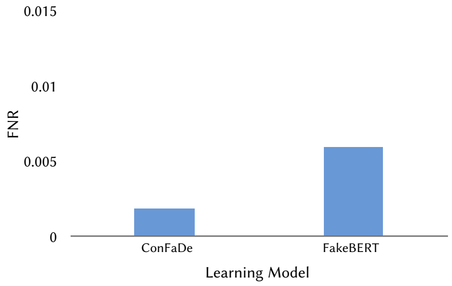


Fig.12. FNR comparison with FakeBERT.

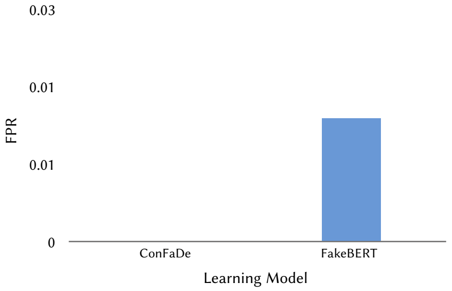


Fig. 13. FPR comparison with FakeBERT.

On comparing both the models on the parameters of FPR and FNR, it can be observed that ConFaDe model has a lower False Negative Rate than FakeBERT and a zero False Positive Rate. It is indicative of the superior performance of the model.

Compared with other existing models that have worked on real-world fake news data, including machine learning models, CNN based models, vanilla LSTM based models, and Hybrid architecture models, the proposed model performs better. The comparison is reported in Table XV.

TABLE XV. Comparison With Existing Works on Real-World Fake News Datasets [36]

Paper	Accuracy Reported (in %)	Technique/Name
(Ghanem et al., 2019)	48.80	SVM,RF,NB, DNN
(Singh et al., 2017)	87.00	SVM
(Ahmed et al., 2017)	89.00	LR- unigram model
(Ruchansky et al., 2017)	89.20	CSI Model
(Ahmed et al., 2017)	92.00	LSVM model
(Liu & Wu, 2018)	92.10	RNN+CNN
(O'Brien et al., 2018)	93.50	Word2Vec+Deep Learning
(Kaliyar et al., 2021)	98.90	FakeBERT(BERT+CNN)
This Paper	**99.90**	**ConFaDe(ELECTRA+LSTM)**

Different models for fake news prediction have been tried in previous works. As is evident from Table XV, machine learning models tend to have low accuracy, owing to the dependence on manual feature engineering. As the deep learning models are used for fake news detection, the accuracy increases sharply. The area of improvement for the detection of fake news in natural language processing condenses to the task of finding an efficient representation of text. Lately, the use of transformers in downstream tasks has been prominently explored. Transformer architectures like BERT are pre-trained on large corpora and afterward used for various downstream tasks, including fake news detection [36].

The central theme of our study is to create efficient as well as accurate architecture for fake news detection. The concerns about the cost and carbon emissions of huge NLP models are reasonably valid, but overlooked in current research. Bender et al. [77] delve into the concerns regarding the environmental and financial costs of NLP models. On training the model once, BERT-base with 110M parameters has an energy consumption of 1507 kWh and carbon emission of 1438 lbs, with a cost of USD 3751-12571 [78]. This study tries to address the problem of fake news through a model that is computationally less expensive, innately simpler, and sustainable in the longer run. The proposed system utilizes a lighter version of ELECTRA, which is about ¼ of the size of BERT, moving a step forward towards Green AI. Though we did not measure the power consumption and carbon emission of the ConFaDe model, it only has 14.9 M parameters [23] (95.1M less parameters than BERT base), and intuitively, its carbon emission and energy consumption will be far less than BERT based models.

The proposed model performs well on a benchmark dataset, outperforming the existing state-of-the-art model in terms of accuracy and efficiency. Although there are many studies related to fake news detection through natural language processing, these studies have not used transformer architecture trained on replaced token detection task to capture context. Previous studies on fake news detection have resulted in creating complex deep learning architectures, adding more layers, and making the process more complex than concentrating on simple solutions. This study focuses on designing simple and efficient system for fake news detection which achieves best results while consuming fewer resources.

Although this study outperforms the existing state-of-the-art model on the same task, there are some limitations. The model being pre-trained on English corpus, is limited to the English language. Fake news detection models need to be developed for multi-lingual fake news where a single news item on internet or social media contains different languages. This study only considers text of the fake news, and as such, multimodal fake news is not covered in this. With respect to the embedding size, fake news detection for long texts is not possible for this model. But, as headlines are an effective cue in detecting the fake news, this model can be applied to the headlines of such lengthy articles. For tackling fake news in different languages like Hindi, Urdu, Arabic, Tamil etc., separate architectures pre-trained for the task are needed.

The technique proposed in the current paper can find its utility in the fake news detection on internet. After creating a deployable model based on the techniques proposed, it can be used to detect fake news on internet portals, microblogs, and social media. Going by the main idea in the proposed technique, small applications for less capable hardware can be developed and deployed. The effective application of the technique proposed may lead to early detection of fake news and thus alleviate the harms caused by the fake news.

VII. Conclusion and Future Scope

In this study, we try to tackle the problem of fake news detection on social media through transformer based contextualized word embeddings. We also conduct experiments with various word embeddings and deep learning models to evaluate the efficiency of each embedding model. We utilize a version of BERT based model-ELECTRA Small++, which differs from original BERT model in the pre-training task and is lighter as far as training and resource consumption is concerned. We generate the word embeddings for LSTM based architecture using the same model. Our word-embedding model is pre-trained on a large English corpus. The results are suggestive of the efficient performance of our model, ConFaDe, over the current state-of-the-art model FakeBERT with the same real-world fake news dataset. We also compare the performance of our model with the FakeBERT model through FPR, FNR and F1 macro and it performs very well on the task with an accuracy of 99.9% and F1 macro of 99.9%.

The model can easily be applied to the English language, as the pre-training task is done on the English corpus. However, for multi-lingual or resource-scarce languages like Urdu and Hindi, we first need to pre-train our transformer model on a corpus of the language. This type of model is best suited for online micro-blogging sites like Twitter and other social media as the sentence size is limited in these platforms.

As a future task, we can incorporate different social and psychological theories and approach the problem with a more data-centric approach towards data-scarce languages to develop a model, which can efficiently tackle the problem of fake news. With a more multi-disciplinary approach, the problem of fake news detection can also be tackled on multiple fronts- from users to network, and the interaction thereof. Another front of work is detecting the fake news in different forms-like picture, text, and video (multimodal). Efficient multimodal fake news detection may be achieved by exploring ensemble of efficient models.

References

[1] X. Zhou and R. Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities," ACM Computing Surveys, vol. 53, no. 5, pp. 1–40, 2020, https://doi.org/10.1145/3395046.

[2] J. Golbeck et al., "Fake news vs satire: A dataset and analysis," in WebSci 2018 - Proceedings of the 10th ACM Conference on Web Science, pp. 17–21, 2018, https://doi.org/10.1145/3201064.3201100.

[3] X. Zhou, R. Zafarani, K. Shu, and H. Liu, "Fake News: Fundamental Theories, Detection Strategies and Challenges," In *Proceedings of the twelfth ACM international conference on web search and data mining*, pp. 836-837, 2019, https://doi.org/10.1145/3289600.3291382.

[4] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," vol. 19, no.1, pp.22-36, 2017, https://doi.org/10.1145/3137597.3137600.

[5] X. Zhou and R. Zafarani, "Fake News: A Survey of Research, Detection Methods, and Opportunities," ACM Computing Surveys, vol. 53, no.5, pp 1–40, 2018, https://doi.org/10.1145/3395046.

[6] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," Science, vol. 359, no. 6380, pp. 1146–1151, 2018, https://doi.org/10.1126/science.aap9559.

[7] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," Journal of Economic Perspectives, vol. 31, no. 2, pp. 211–236, 2017, https://doi.org/10.1257/jep.31.2.211.

[8] M. H. Alkawaz and S. A. Khan, "Use of Fake News and Social Media by Main Stream News Channels of India," in Proceedings - 2020 6th IEEE International Colloquium on Signal Processing & Its Applications, pp. 93–97, 2020, https://doi.org/ 10.1109/CSPA48992.2020.9068673.

[9] S. Lewandowsky, U. K. H. Ecker, C. M. Seifert, N. Schwarz, and J. Cook, "Misinformation and Its Correction: Continued Influence and Successful Debiasing," Psychological Science in the Public Interest, Supplement, vol. 13, no. 3, pp. 106–131, 2012, https://doi.org/10.1177/15291006124510.

[10] N. Walter and R. Tukachinsky, "A Meta-Analytic Examination of the Continued Influence of Misinformation in the Face of Correction: How Powerful Is It, Why Does It Happen, and How to Stop It?," Communication Research, vol. 47, no. 2, pp. 155–177, 2020, https://doi.org/10.1177/0093650219854600.

[11] Y. Tsfati, H. G. Boomgaarden, J. Strömbäck, R. Vliegenthart, A. Damstra, and E. Lindgren, "Causes and consequences of mainstream media dissemination of fake news: literature review and synthesis," vol. 44, no. 2, pp. 157–173, 2020, https://doi.org/10.1080/23808985.2020.1759443.

[12] R. Zafarani, X. Zhou, K. Shu, and H. Liu, "Fake news research: Theories, detection strategies, and open problems," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 3207–3208, 2019, https://doi.org/10.1145/3292500.3332287.

[13] T.A. Shaikh, W.A. Mir, I. Mohammad, and R. Ali, "An Intelligent Healthcare System for Automated Alzheimer's Disease Prediction and Personalized Care," International Journal of Next-Generation Computing, vol. 12, no. 2, pp.240-253, 2021, https://doi.org/10.47164/ijngc.v12i2.196.

[14] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in WSDM 2019 - Proceedings of the

12th ACM International Conference on Web Search and Data Mining, pp. 312–320, 2019, https://doi.org/10.1145/3289600.3290994.

[15] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," in *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55-75, 2018, https://doi.org/10.1109/MCI.2018.2840738.

[16] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A hybrid deep model for fake news detection," in International Conference on Information and Knowledge Management, Proceedings, Nov. 2017, pp. 797–806, 2017, https://doi.org/10.1145/3132847.3132877.

[17] C. Song, K. Shu, and B. Wu, "Temporally evolving graph neural network for fake news detection," Information Processing & Management, vol. 58, no. 6, 2021, https://doi.org/10.1016/j.ipm.2021.102712.

[18] L. Weng, F. Menczer, and Y. Y. Ahn, "Virality prediction and community structure in social networks," Scientific Reports, vol. 3, no. 1, pp. 1–6, 2013, https://doi.org/10.1038/srep02522.

[19] D. Centola, "The spread of behavior in an online social network experiment," Science, vol. 329, no. 5996, pp. 1194–1197, 2010, https://doi.org/10.1126/science.1185231.

[20] R. Oshikawa, J. Qian, and W. Y. Wang, "A survey on natural language processing for fake news detection," in LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings, European Language Resources Association (ELRA), 2020, pp. 6086–6093.

[21] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp. 1532–1543, 2014, http://dx.doi.org/10.3115/v1/D14-1162.

[22] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, pp. 4171–4186, 2019, https://doi.org/10.18653/v1/n19-1423.

[23] K. Clark, M.-T. Luong, Q. v. Le, and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," *arXiv preprint arXiv:2003.10555, 2020,* https://doi.org/10.48550/arXiv.2003.10555.

[24] A. Bondielli, F. Marcelloni "A survey on fake news and rumour detection techniques," Information Sciences, vol. 497, pp. 38–55, 2019, https://doi.org/10.1016/j.ins.2019.05.035.

[25] V. K. Singh, R. Dasgupta, D. Sonagra, K. Raman, and I. Ghosh, "Automated Fake News Detection Using Linguistic Analysis and Machine Learning," In International conference on social computing, behavioral-cultural modeling, & prediction and behavior representation in modeling and simulation (SBP-BRiMS), pp. 1-3, 2017, https://doi.org/10.13140/RG.2.2.16825.67687.

[26] S. P. Yadav, "Emotion recognition model based on facial expressions," Multimedia Tools and Applications, vol. 80, no. 17, pp. 26357–26379, 2021, https://doi.org/10.1007/s11042-021-10962-5.

[27] S. P. Yadav, "Vision-based detection, tracking, and classification of vehicles," IEIE Transactions on Smart Processing and Computing, vol. 9, no. 6, pp. 427–434, 2020, https://doi.org/10.5573/IEIESPC.2020.9.6.427.

[28] V. L. Rubin, Y. Chen, and N. J. Conroy, "Deception Detection for News: Three Types of Fake News," in Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community, pp.1–4, 2015, https://doi.org/10.1002/pra2.2015.145052010083.

[29] C. Castillo, M. Mendoza, and B. Poblete, "Predicting information credibility in time-sensitive social media," Internet Research, vol. 23, no. 5, pp. 560–588, 2013, https://doi.org/10.1108/IntR-05-2012-0095.

[30] H. Ahmed, I. Traore, and S. Saad, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques," in Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol. 10618, pp. 127–138, 2017, https://doi.org/10.1007/978-3-319-69155-8_9.

[31] A. Giachanou, G. Zhang, and P. Rosso, "Multimodal fake news detection with textual, visual and semantic information," In Proceedings of Text, Speech, and Dialogue: 23rd International Conference, Springer-Verlag, Berlin, Heidelberg, pp. 30–38, 2020, https://doi.org/10.1007/978-3-030-58323-1_3.

[32] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru and S. Satoh, "SpotFake: A Multimodal Framework for Fake News Detection," 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), Singapore, pp. 39-47, 2019, https://doi.org/10.1109/BigMM.2019.00-44.

[33] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 1st International Conference on Learning Representations, Workshop Track Proceedings, USA, 2013.

[34] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016, https://doi.org/10.1109/CVPR.2016.90.

[35] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261-2269, 2017, https://doi.org/10.1109/CVPR.2017.243.

[36] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," Multimedia Tools and Applications, vol. 80, no. 8, pp. 11765–11788, 2021, https://doi.org/10.1007.

[37] K. Cho, B.van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1724–1734, 2014, https://doi.org/10.3115/v1/d14-1179.

[38] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink and J. Schmidhuber, "LSTM: A Search Space Odyssey," in IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 10, pp. 2222-2232, Oct. 2017, https://doi.org/10.1109/TNNLS.2016.2582924.

[39] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997, https://doi.org/10.1162/neco.1997.9.8.1735.

[40] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," in IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673-2681, 1997, https://doi.org/10.1109/78.650093.

[41] S. Girgis, E. Amer and M. Gadallah, "Deep Learning Algorithms for Detecting Fake News in Online Text," 2018 13th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, 2018, pp. 93-97, https://doi.org/10.1109/ICCES.2018.8639198.

[42] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," In Proceedings of NAACL-HLT, pp. 4171-4186, 2019, https://doi.org/10.18653/v1/n19-1423.

[43] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017, pp. 5999–6009.

[44] S. A. Alkhodair, B. C. M. Fung, S. H. H. Ding, W. K. Cheung, and S. C. Huang, "Detecting High-Engaging Breaking News Rumors in Social Media," ACM Transactions on Management Information Systems, vol. 12, no. 1, pp. 1-16, 2021, https://doi.org/10.1145/3416703.

[45] T. Bian et al., "Rumor detection on social media with bi-directional graph convolutional networks," in Proceedings of the AAAI conference on artificial intelligence vol. 34, no. 01, pp. 549-556, 2020.

[46] J. Shin, L. Jian, K. Driscoll, and F. Bar, "The diffusion of misinformation on social media: Temporal pattern, message, and source," Computers in Human Behavior, vol. 83, pp. 278–287, 2018, https://doi.org/10.1016/j.chb.2018.02.008.

[47] X. Zhou and R. Zafarani, "Network-based Fake News Detection: A Pattern-driven Approach," ACM SIGKDD Explorations Newsletter, vol. 21, no. 2, pp. 48–60, 2019, https://doi.org/10.1145/3373464.3373473.

[48] C. Song, C. Yang, H. Chen, C. Tu, Z. Liu and M. Sun, "CED: Credible Early Detection of Social Media Rumors," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 8, pp. 3035-3047, 2021, https://doi.org/10.1109/TKDE.2019.2961675.

[49] H. G. Oliveira, T. Sousa, A. Alves, "Assessing Lexical-Semantic Regularities in Portuguese Word Embeddings," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 5, pp. 34-46, 2021, https://doi.org/10.9781/ijimai.2021.02.006.

[50] S. P. Yadav, S. Zaidi, A. Mishra, and V. Yadav, "Survey on Machine Learning in Speech Emotion Recognition and Vision Systems Using a Recurrent Neural Network (RNN)," Archives of Computational Methods in Engineering, vol. 29, no. 3, pp. 1753–1770, 2021, https://doi.org/10.1007/s11831-021-09647-x.

[51] J. Ma et al., "Detecting rumors from microblogs with recurrent neural networks," In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, pp. 3818–3824, 2016, https://dl.acm.org/doi/10.5555/3061053.3061153.

[52] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in Proceedings of the 31st International Conference on International Conference on Machine Learning, 2014, pp. 1188–1196. https://dl.acm.org/doi/10.5555/3044805.3045025.

[53] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," In Proceedings of International conference on learning representations, 2017.

[54] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," In Proceedings of the 26th International Conference on Neural Information Processing Systems, vol. 2, pp. 3111–3119, 2013.

[55] A. Agarwal, M. Mittal, A. Pathak, L. M. Goyal, "Fake News Detection Using a Blend of Neural Networks: An Application of Deep Learning," S.N. Computer Science, vol. 1, no. 3, 2020 https://doi.org/10.1007/s42979-020-00165-4.

[56] G. K. W. Huang and J. C. Lee, "Hyperpartisan News and Articles Detection Using BERT and ELMo," 2019 International Conference on Computer and Drone Applications (IConDA), Kuching, Malaysia, pp. 29-32, 2019, https://doi.org/10.1109/IConDA47345.2019.9034917.

[57] R. M. Silva, R. L. S. Santos, T. A. Almeida, and T. A. S. Pardo, "Towards automatically filtering fake news in Portuguese," Expert Systems with Applications, vol. 146, pp. 113199, 2020, https://doi.org/10.1016/j.eswa.2020.113199.

[58] M. Samadi, M. Mousavian, and S. Momtazi, "Deep contextualized text representation and learning for fake news detection," Information Processing & Management, vol. 58, no. 6, pp. 102723, 2021, https://doi.org/10.1016/j.ipm.2021.102723.

[59] M. Hoang, O. A. Bihorac, and J. Rouces, "Aspect-Based Sentiment Analysis using BERT," In Proceedings of the 22nd Nordic conference on computational linguistics, pp. 187-196, 2019.

[60] A. Alessa, M. Faezipour and Z. Alhassan, "Text Classification of Flu-Related Tweets Using FastText with Sentiment and Keyword Features," In Proceedings of IEEE International Conference on Healthcare Informatics, pp. 366-367, 2018, https://doi.org/10.1109/ICHI.2018.00058.

[61] B. Büyüköz, A. Hürriyetoğlu, and A. Özgür, "Analyzing ELMo and DistilBERT on Socio-political News Classification," In Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020, pp. 9-18, 2020.

[62] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner and G. Monfardini, "The Graph Neural Network Model," in IEEE Transactions on Neural Networks, vol. 20, no. 1, pp. 61-80, 2009, https://doi.org/10.1109/TNN.2008.2005605.

[63] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar and M. S. Rahman, "A Comprehensive Review on Fake News Detection with Deep Learning," in IEEE Access, vol. 9, pp. 156151-156170, 2021, https://doi.org/10.1109/ACCESS.2021.3129329.

[64] B. Khoo, R. C. W. Phan, and C. H. Lim, "Deepfake attribution: On the source identification of artificially generated images," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 12, no. 3, p. e1438, 2022, https://doi.org/10.1002/widm.1438.

[65] J. F. Low, B. C. M. Fung, F. Iqbal, and S. C. Huang, "Distinguishing between fake news and satire with transformers," Expert Systems with Applications, vol. 187, pp. 115824, 2022, https://doi.org/10.1016/j.eswa.2021.115824.

[66] J. A. Reshi and R. Ali, "Rumor proliferation and detection in Social Media: A Review," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019, pp. 1156-1160, https://doi.org/10.1109/ICACCS.2019.8728321.

[67] J. Lies, "Marketing Intelligence: Boom or Bust of Service Marketing?," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 7, no. 7, pp. 115-124, 2022, https://doi.org/10.9781/ijimai.2022.10.001.

[68] S. A. Alameri and M. Mohd, "Comparison of Fake News Detection using Machine Learning and Deep Learning Techniques," In Proceedings of 2021 3rd International Cyber Resilience Conference (CRC), pp. 1-6, 2021, https://doi.org/10.1109/CRC50527.2021.9392458.

[69] M. Z. Khan and O. H. Alhazmi, "Study and analysis of unreliable news based on content acquired using ensemble learning (prevalence of fake news on social media)," International Journal of Systems Assurance Engineering and Management, vol. 11, no. 2, pp. 145–153, 2020, https://doi.org/10.1007/s13198-020-01016-4.

[70] V.L. Rubin, "Artificially Intelligent Solutions: Detection, Debunking, and Fact-Checking," In Misinformation and Disinformation, pp. 207-263, 2022, https://doi.org/10.1007/978-3-030-95656-1_7

[71] D. P. Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization," arXiv preprint arXiv:1412.6980, 2014.

[72] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing and Management: an International Journal, vol. 45, no. 4, pp. 427–437, 2009, https://doi.org/10.1016/j.ipm.2009.03.002.

[73] Z. C. Lipton, C. Elkan, and B. Naryanaswamy, "Optimal thresholding of classifiers to maximize F1 measure," in Lecture Notes in Computer Science, vol. 8725, pp. 225–239, 2014, https://doi.org/10.1007/978-3-662-44851-9_15.

[74] B. Ghanem, P. Rosso, and F. Rangel, "Stance Detection in Fake News A Combined Feature Representation," in Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pp. 66–71, 2018, https://doi.org/10.18653/v1/W18-5510.

[75] Y. Liu and Y.-F. B. Wu, "Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018, https://doi.org/10.1609/aaai.v32i1.11268

[76] N. O'Brien, S. Latessa, G. Evangelopoulos, and X. Boix, "The Language of Fake News: Opening the Black-Box of Deep Learning Based Detectors," In Proceedings of workshop on "A.I. for Social Good," 32nd Conference on Neural Information Processing Systems, 2018.

[77] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623, 2021, https://doi.org/10.1145/3442188.3445922.

[78] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3645–3650, 2019, https://doi.org/10.18653/v1/P19-1355.

Junaid Ali Reshi

Junaid Ali Reshi received his B.Tech in Computer Science and Engineering (2015) from University of Kashmir and M. Tech in Computer Science and Technology (2017) from Central University of Punjab, India. He is currently a doctoral research scholar in the department of Computer Engineering, Aligarh Muslim University, India. He has served as reviewer for many papers in WoS indexed journals and top conferences, and is actively involved in research. His areas of interest are Social Network analysis, Natural Language processing, and Computational Social science.

Rashid Ali

Prof. Rashid Ali obtained his B.Tech. and M.Tech. from Aligarh Muslim University, India in 1999 and 2001 respectively. He obtained his PhD in Computer Engineering in February 2010 from Aligarh Muslim University, India. He is currently serving as full Professor in the department of Computer Engineering, Aligarh Muslim University. He is also serving as the Coordinator, Interdisciplinary Center for Artificial Intelligence, Aligarh Muslim University. Apart from being a member of various international societies, he is also a senior member of IEEE. He has authored more than 125 papers in various International Journals and conferences of repute. He has also chaired sessions at some International conferences. He reviews articles for some of the reputed International Journals and conferences. He has supervised more than 25 M.Tech Dissertation and 6 PhD Thesis. His research interests include Web Searching, Web Mining, Soft computing Techniques (Rough-Set, Artificial Neural Networks, fuzzy logic etc.), Recommender Systems, and Online Social Network Analysis.

Graffiti Identification System Using Low-Cost Sensors

Miguel García García[1]*, Angélica González Arrieta[1,2]*, Sara Rodríguez González[1,2]*, Sergio Márquez-Sánchez[2], Carlos Fernando Da Silva Ramos[3]

[1] Dept. of Computer Science and Automation, Faculty of Science, University of Salamanca, Plaza de los Caídos s/n, 37008, Salamanca (Spain)
[2] BISITE Research Group, University of Salamanca, Edificio I+D+i, Calle Espejo 2, Salamanca 37007, Salamanca (Spain)
[3] Institute of Engineering, Polytechnic of Porto, Rua Dr. António Bernardino de Almeida 431, 4200-072, Porto (Portugal)

* Corresponding author. miguel@restnova.net (M. García García), angelica@usal.es (A. González Arrieta), srg@usal.es (S. Rodríguez González).

Received 28 April 2022 | Accepted 30 January 2023 | Early Access 8 May 2023

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

This article introduces the possibility of studying graffiti using a colorimeter developed with Arduino hardware technology according to the Do It Yourself (DIY) philosophy. Through the obtained Red Green Blue (RGB) data it is intended to study and compare the information extracted from each of the graffiti present on different walls. The same color can be found in different parts of a single graffiti, but also in other graffiti that could a priori be of different authorship. Nevertheless, graffiti may be related, and it may be possible to group graffiti artists and "gangs" that work together. The methodology followed for the construction of the colorimeter and its real application in a practical case are described in four case studies. The case studies describe how graffiti were identified and recognized and they provide a comparison of the collected color samples. The results show the added value of the colorimeter in the graffiti recognition process, demonstrating its usefulness on a functional level. Finally, the contributions of this research are outlined, and an analysis is carried out of the changes to be made to the proposed method in the future, for improved graffiti color identification.

Keywords

Colorimeter, Crew, Graffiti, Image Color Analysis, Low-Cost Sensors, Tag.

DOI: 10.9781/ijimai.2023.05.001

I. Introduction

The identification and classification of graffiti is a discipline of interest, not only to authors belonging to this urban tribe or graffiti writers, but also to institutions interested in controlling and reducing the presence of this type of vandalism. There are different classification systems proposed from the very moment in which a graffiti is created [1] and stylistic or calligraphic systems are used in its identification [2]. Just as it is possible for an expert calligrapher to analyze signatures, it is also possible to classify graffiti by identifying its characteristics.

This study aims to assess the use of a low-cost TCS34725 color sensor and a colorimeter, to automate the identification of graffiti [3], assuming that this tool is capable of improving graffiti characterization accuracy, as this factor had not been taken into account when identifying it (from the state of the art, it could be verified that images may be analyzed using this colorimeter, however, their chromatic “reliability” had not been considered).

The function of a colorimeter is to accurately determine the color applied to a surface. There are different types of colors on the market, depending on the intended use and the default color pattern in different numerical parameters (RGB, CMYK, etc.). The applications of this device include determining a color on a wall to produce another with the same characteristics or determining the ripening point of different fruits. Currently there are works which have used those devices, some in the field of health [4] and others in the analysis and interpretation of digital images [5], [6]. There are previous works on color analysis and the construction of low-cost sensors [7], graffiti documentation methods [8] and artificial vision [9]. Regarding the analysis of graffiti and paintings, there are studies regarding their chemical analysis [10], the prevention of graffiti [11] or the analysis of color in rock art [12]. In this case, it has been decided to build the colorimeter to understand the entire process and thus determine what needs to be taken into account for its improvement in the future.

The colorimeter is composed of a red, green and blue light-emitting diode (RGB LED) [13], a focusing lens, a 3D-printed cuvette holder, and a light-sensitive photodiode detector. The processing technology of the Arduino microcontroller has been used. This device simulates the response of the standard observer and can match most colors in the visible spectrum. Using an internal light source, the colorimeter illuminates the sample surface; As light reflects off the device, it

Please cite this article as:
M. García García, A. González Arrieta, S. Rodríguez González, S. Márquez-Sánchez, C. F. Da Silva Ramos, “Graffiti Identification System Using Low-Cost Sensors”, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 6, pp. 51-60, 2024, http://dx.doi.org/10.9781/ijimai.2023.05.001

passes through three filters: red, green, and blue. These filters distill the values of the three stimuli (RGB) that match the way our eyes see color according to its intensity, as illustrated in Fig. 1 and Fig. 2 [14].

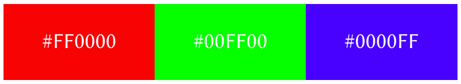


Fig. 1. RGB color code.

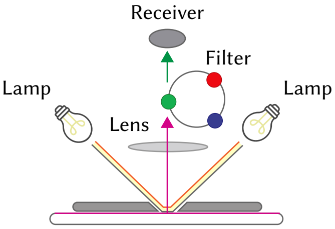


Fig. 2. Colorimeter operation.

Arduino is an open-source platform for the creation of electronics, based on free hardware and software. It is flexible and easy to use for creators and developers. This platform enables the creation of different types of "small computers" [15] on a single board, that can be put to different types of use by communities of creators, individual users and developers. As the hardware is free it offers a basis on which a person or company can create their own boards, which may be different from each other, but equally functional. Moreover, because the software is free, it offers the Arduino IDE platform (Integrated Development Environment) [16]; a programming environment with which anyone can create applications for Arduino boards [17], so that they can be given all kinds of utilities. Their code is made accessible to all, so that anyone can use it and modify it [18].

Therefore, Arduino is based on a board with a microcontroller as hardware [19], and a development environment called IDE, which is an open and free software that facilitates the use of electronics and microcontrollers in multidisciplinary projects [20].

On this platform, a colorimeter has been built using the TCS34725 digital color sensor [21], [22]. This is intended to obtain a sufficiently reliable data relationship based on the different color samples.

The research described in this article has been carried out with the aim of achieving the following objectives:

- Achieve the functionality of a colorimeter built with Arduino.
- Check that the data obtained with the colorimeter are reliable.
- Evaluate the colorimeter as a graffiti identification tool.

The rest of the article is structured as follows: Section II analyzes the previously proposed related methods and techniques for color identification. In Section III, the proposed system is described. In Section IV the experimental results are presented and finally, conclusions are drawn regarding the performance of the proposed system and future work is outlined in Section V.

II. Preliminaries

The colorimeter is a device with special sensitivity which enables it to detect different color ranges. For example, it is used to determine the ripening point of a fruit [23] or the color of a food according to the consumer's tolerance level (studies with meat products [24] and dairy products [25]). It is also used to detect a color that needs to be reproduced in the automotive industry or in interior decoration [26]. In this specific case, the objective is to use it in the identification of graffiti from the obtained data by reading the colors through the sensor.

The TCS34725 (Fig. 3. Sensor TCS34725.) is a sensor that is capable of calculating the value of the different components that make up the color through Inter-Integrated Circuit (I2C) communication, the sensor sends the data on the level of red, green, blue light present in a color. The communications make it very easy to integrate it in any device controlled by a main control unit (MCU), as it comes with a development board for the implementation of integrated circuits.

This sensor's features include RGB LED backlight control, light color temperature measurement, ambient light detection for screen backlight control, fluid and gas analysis, and verification and classification of the color of the products. For this reason, it can be found in televisions, smartphones, tablets, computers, monitors, printers, in medical devices, in signaling devices or in industrial automation.

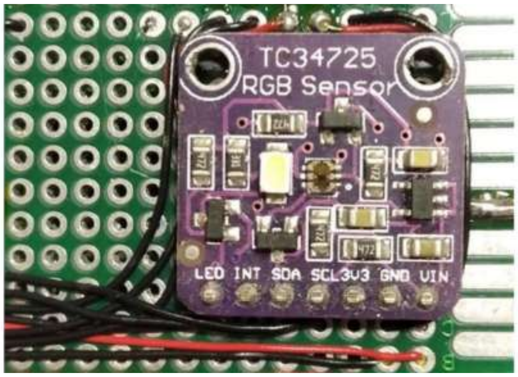


Fig. 3. Sensor TCS34725.

The operation of the TCS34725 [27] consists of an array of 3x4 reverse biased photodiodes which produce a small electric current when excited by light of a certain wavelength. In addition, they have an infrared filter (IR) to minimize the response to this range of the spectrum (Fig. 4. Representation of the photodiode array in the package. Response graph of photodiodes as a function of wavelength (TCS34725 Datasheet).

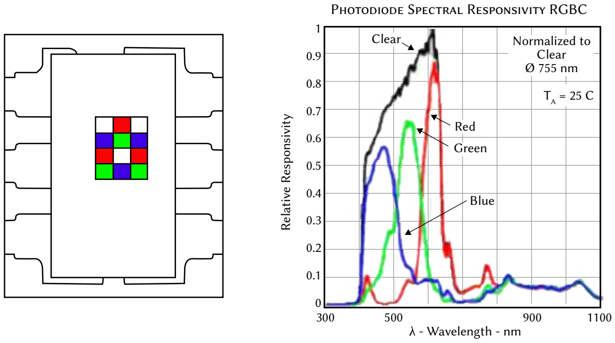


Fig. 4. Representation of the photodiode array in the package. Response graph of photodiodes as a function of wavelength (TCS34725 Datasheet).

The small current generated by the photodiode causes a change in its voltage when excitation occurs. The variation in voltage is measured with an Analog-Digital Converter (ADC) (Fig. 5. Sensor functional block diagram (TCS34725 Datasheet).

The ADC converters integrate the measurement of the photodiodes [28]. These measurements are transferred to the internal registers of the TCS34725 [29], which incorporate a double buffer to ensure data integrity. The state of the sensor and the energy is controlled by an internal state machine, which controls all the functions of the TCS34725 [30]. The sensor has a dynamic range of 3,800,000:1 with adjustable integration time and gain, making it suitable for use behind

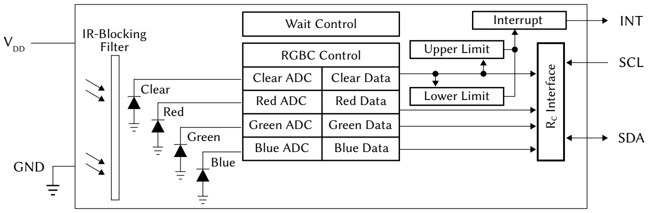


Fig. 5. Sensor functional block diagram (TCS34725 Datasheet).

tinted glass or fabric. The programmable interrupt pin allows for interruptions in the level of light when preset values are exceeded, thus reducing microprocessor overload. The TCS34725 sensor has a low power standby state to reduce average power consumption:

- Low consumption: 2.5µA in sleep mode.
- 65µA in standby mode, with programmable standby time from 2.4ms to 7 seconds.

Taking into account the existence of works in other fields [31] related to the identification of mural paintings [32] or the identification of graffiti on roads [33], we consider that the use of the colorimeter can be assessed in relation to different previously selected graffiti on a public street wall. The section that follows describes the construction of the colorimeter which has made it possible to capture colorimetric values related to the graffiti in question [34].

III. Graffiti Identification System

This section describes the developed system, both its functional and electronic description.

A. Functional Description

The system can be divided into the following functional steps (Fig. 7):

- Step 1. Selection of the area graffiti image collection. In this step, the area to be measured was selected, taking into account the graffiti image to be studied.
- Step 2. Image reading through RGB levels by the TCS34725 sensor. The TCS3472 Light-to-Digital Converter contains a 3x4 photodiode array consisting of red, green, blue, and clear (unfiltered) IR-cut filter-coated photodiodes. These determine the color when excited inversely with a certain wavelength, generating a small current that is transformed into a signal thanks to the use of four analog-digital converters integrated in the photodiodes, which simultaneously convert the amplified currents of the photodiodes into a 16-bit digital value. At the end of a conversion cycle, the results are transferred to the data registers and double buffering ensures data integrity. For internal timing control and low power wait state, a state machine is used.
- Step 3. Image processing by means of image fragmentation. The RGBC engine (Fig. 6. RGBC operation.) contains the gain control and four analog-digital converters integrated in the photodiodes. The integration of all four channels occurs simultaneously, and once the conversion cycle is complete, the results are transferred to the color data registers, doing what is called a channel count. An important issue to consider is the integration time, as it affects both the resolution and sensitivity of the RGBC readout. The communication of the TCS3472 data is carried out through a transfer of up to 400 kHz, two-wire of the I2C standard, which facilitates direct connection with microcontrollers and embedded processors.
- Step 4. Comparison of values with the cathode and corresponding LED lighting. Data reads are double buffered to ensure that no invalid data is read during the transfer. The red, green, blue and clear (unfiltered) read data is stored as 16-bit values. To ensure that this data has been read correctly, a two-byte I2C read is performed, with single word bit reading set by the command register. If both values are equal, the transfer is verified, and the device automatically passes to the next state according to the configured state machine.

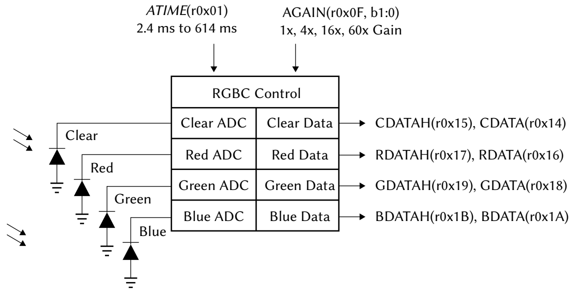


Fig. 6. RGBC operation.

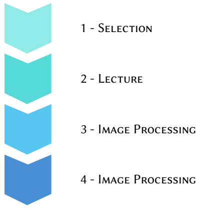


Fig. 7. Funcional description.

B. Sensor Description

Shooting is affected by different values. In the case of the TCS34725 sensor, it is expected to be able to obtain reliable color samples for the comparison of some graffiti with others, making it possible to establish a link between them in terms of color.

The purpose of the device is to measure color and display the levels of each RGB component on a liquid crystal display (LCD). Additionally, an RGB LED has been added to display the measured color. The scheme of the device is shown in Fig. 8.

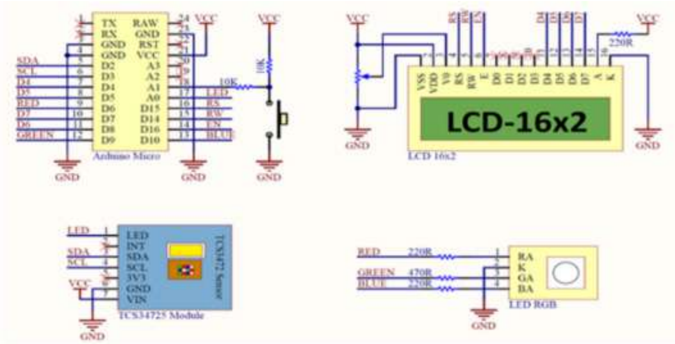


Fig. 8. Schematic of the electronic assembly.

The main controller is an Arduino Micro development board [35]. The device is powered by the micro-USB connector integrated in the Arduino board. Through the VCC pin of the Arduino, 5V power is provided to the rest of the components of the circuit.

Connected to pin A1 is a push button. Pressing it initiates the color reading. To do this, the LED integrated in the TCS34725 module, which is connected to pin A0 of the Arduino, lights up. The connection of the TCS34725 module is simple, it is done through the I2C of the Arduino, pin D2 for the data line and D3 for the clock signal.

Once the reading was done, the different values for RGB were shown on the LCD display. This is a 2-line alphanumeric display with 16 characters per line. The connection to the Arduino was configured for 4-bit parallel communication. Also, on the V0 pin of the display a potentiometer has been added to manually adjust the contrast of the LCD.

Lastly, with the measured RGB values, the outputs of the Arduino are adjusted so that the RGB LED is lit in the same color as has been measured. To do this, a line for the RGB component was used; D6 for red, D9 for green and D10 for blue. The RGB LED is common cathode.

The entire circuit has been mounted on a perforated plate as shown in Fig. 9.

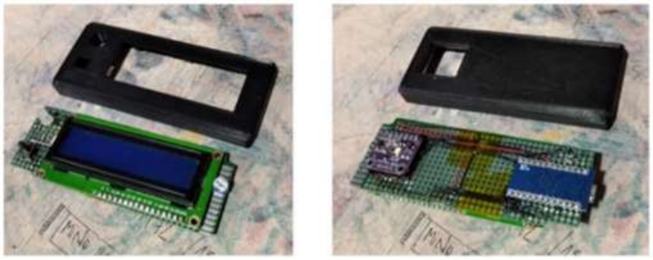


Fig. 9. Colorimeter assembly.

IV. Experimental Results

In the experimental results section, two subsections are presented. The first one presents a preliminary investigation which has been carried out on the abandoned walls located in the vicinity of the Faculty of Fine Arts of the University of Salamanca and presents a measurement and color acquisition experiment using the colorimeter built in the current study. The second subsection presents three case studies demonstrating how the use of this type of methodology is suitable for the automatic characterization and authorship analysis of graffiti.

A. Color Acquisition Preliminary Experiment

The present investigation has been carried out on derelict walls located in the vicinity of the Faculty of Fine Arts of the University of Salamanca, in the vicinity of the city of Salamanca but far enough to house a considerable number of graffiti and tags or signatures (Fig. 10 and Fig. 11).

At first, the photographic documentation of the graffiti to be studied was carried out, as well as its measurement. Following the documentation of the graffiti to be studied, those that had similar colors were chosen for the purposes of this research. Then, the digital color taking process began, using the color selection platform "Adobe Color" [36] and the colorimeter [37], providing the data to obtain the color codes.

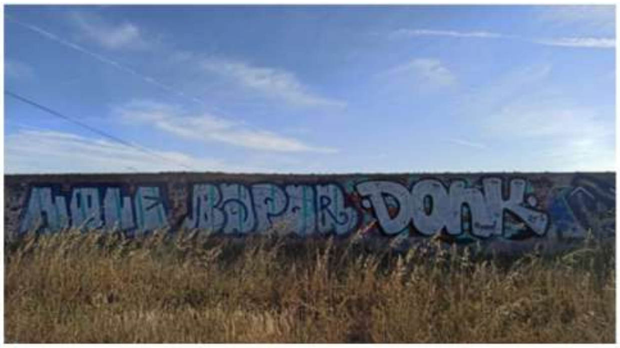


Fig. 10. Graffiti wall. Case study. (Redmi Note 8 Pro, Xiaomi f/1.89, 1/653, ISO100 5.3mm, without Flash).

Three graffiti (Table I) have been analyzed to carry out the preliminary investigation. These are placed continuously and, apparently, use the same colors in their execution. These colors are white/silver, Prussian blue, Ultramarine blue, black and red.

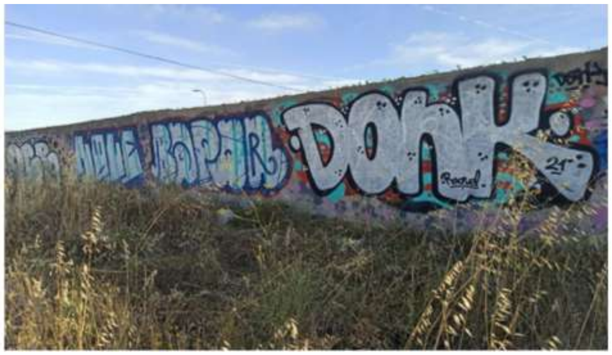


Fig. 11. Lateral view graffiti case study. (Redmi Note 8 Pro, Xiaomi f/1.89, 1/325, ISO102, 5.43mm, without Flash).

TABLE I. Graffiti to Study

GRAFITI 1	WAVE	183 x 420 cm.
GRAFITI 2	BAPOR	180 x 430 cm.
GRAFITI 3	DONK	180 x 425 cm.

We started with the premise that the colorimeter, supported by a TCS34725 sensor, would obtain reliable values. In addition to taking several measurements of the same-colored surface, a color chart (CHROME GUIDE 1650 colors by Valentine) had initially been used as a reference. By taking this step, we intended to make the sampling through the sensor more reliable and thus ensure that the sampling is in a correct RGB measurement range.

Firstly, digital images of the graffiti were taken, and their colors were obtained through the "Adobe Colors" platform, obtaining the results shown in Fig. 12, Fig. 13, Fig. 14, Fig. 15, Fig. 16, Fig. 17 and Fig. 18.

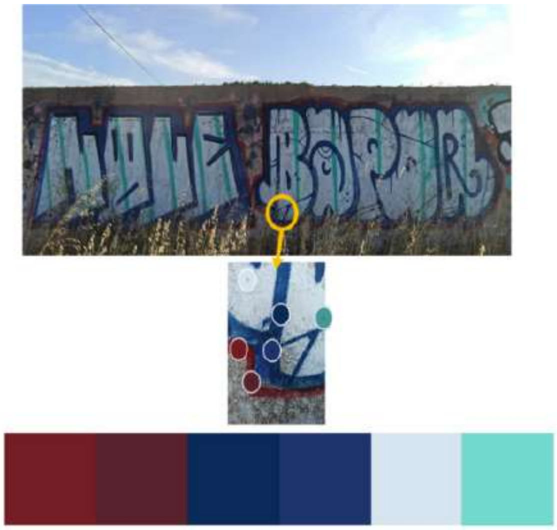


Fig. 12. Processing and comparison of the different parts of graffiti 2 (BAPOR) with the color sensor (Redmi Note 8 Pro, Xiaomi f/1.89, 1/415, ISO100, 5.43mm, without Flash).

Subsequently, the colors were taken with the colorimeter and compared with the color chart. With these two tools, we focused on identifying the different types of blue in both pieces of graffiti for comparison.

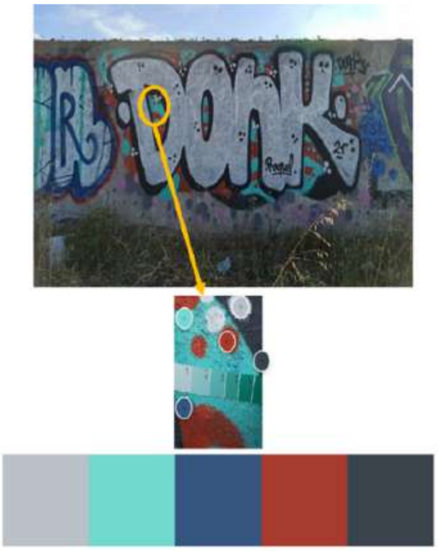
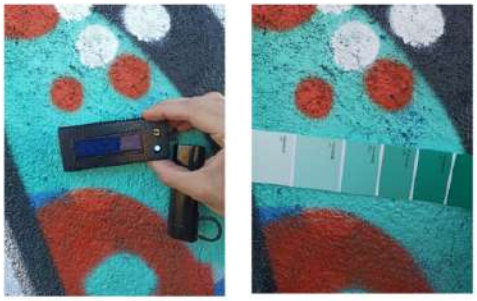
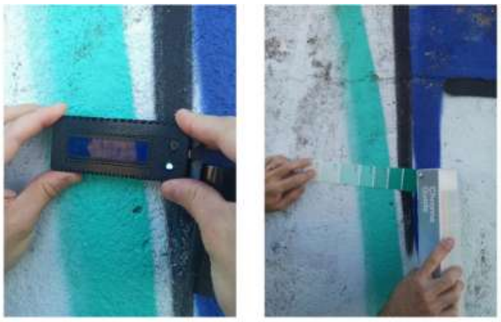


Fig. 13. Different parts of graffiti 3 (DONK). Processing and comparison with the color sensor (Redmi Note 8 Pro, Xiaomi f/1.89, 1/360, ISO100, 5.43mm, without Flash).



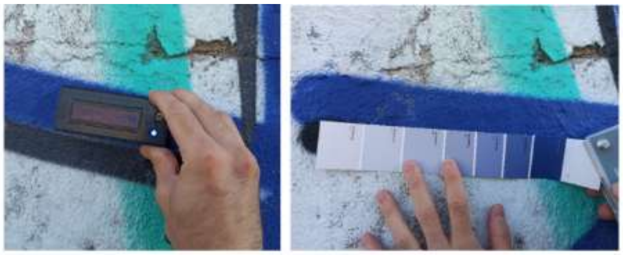
RGB: 48, 105, 89
🟩

Fig. 14. Comparative study of the data obtained with the sensor and the color chart (graffiti 3). (Redmi Note 8 Pro, Xiaomi f/1.89, 1/100, ISO118, 5.43mm, without Flash) (Redmi Note 8 Pro, Xiaomi f/1.89, 1/123, ISO100, 5.43mm, without Flash).



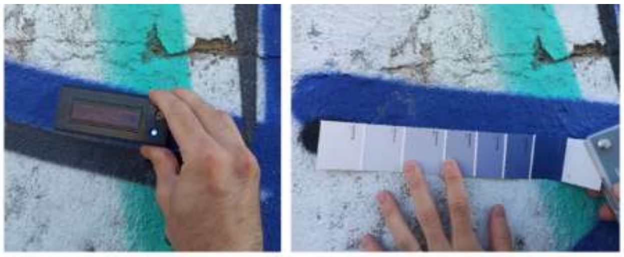
RGB: 39, 110, 85
🟩

Fig. 15. Comparative study of the data obtained with the sensor and the color chart (graffiti 1) (Redmi Note 8 Pro, Xiaomi f/1.89, 1/100, ISO120, 5.43mm, without Flash) (Redmi Note 8 Pro, Xiaomi f/1.89, 1/141, ISO100, 5.43mm, without Flash).



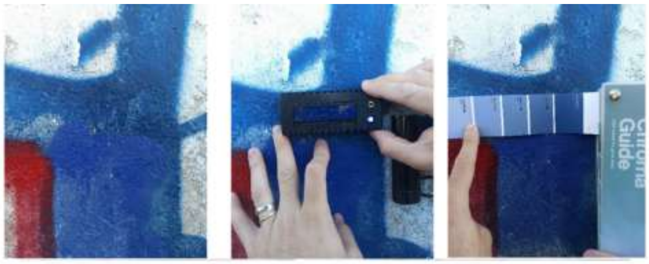
RGB: 30, 71, 136
🟦

Fig. 16. Comparative study of the data obtained with the sensor and the color chart (graffiti 1) (Redmi Note 8 Pro, Xiaomi f/1.89, 1/168, ISO100, 5.43mm, without Flash) (Redmi Note 8 Pro, Xiaomi f/1.89, 1/162, ISO100, 5.43mm, without Flash).



RGB: 35, 74, 172
🟦

Fig. 17. Comparative study of the data obtained with the sensor and the color chart (graffiti 1) (Redmi Note 8 Pro, Xiaomi f/1.89, 1/168, ISO100, 5.43mm, without Flash) (Redmi Note 8 Pro, Xiaomi f/1.89, 1/162, ISO100, 5.43mm, without Flash).



RGB: 48, 84, 125
🟦

Fig. 18. Comparative study of the data obtained with the sensor and the color chart (graffiti 2) (Redmi Note 8 Pro, Xiaomi f/1.89, 1/100, ISO163, 5.43mm, wihout Flash) (Redmi Note 8 Pro, Xiaomi f/1.89, 1/100, ISO179, 5.43mm, without Flash) (Redmi Note 8 Pro, Xiaomi f/1.89, 1/100, ISO141, 5.43mm, without Flash).

In this case, the advantages of using the colorimeter included:

- Simple design and easy handling and mobility (portable).
- Reduced cost of materials.
- Electric charge easy to manage with powerbank-USB.
- Obtaining RGB color data is independent of ambient light.

There are some drawbacks related to colorimeters [9], such as a limited measurement area (2 cm x 2 cm), the restriction that the sample must have a uniform appearance, the location of the sample and the number of readings is significant for representative color evaluation and reflectance properties may be cause of disturbances in color detections.

B. Using Colorimeter to Identify Graffiti Authorship

Once the preliminary study of the colorimeter has been carried out, three case studies are presented in this results subsection. It is demonstrated how the use of this type of methodology is adequate for the automatic characterization and analysis of graffiti authorship. The following case studies have been carried out:

1. A case study on graffiti artists from the same or different crews. The crew, in its most general sense, speaks of a group made up of people who generally work under the direction of the same leader. The first case study presents a study of the graffitist crews: WAVE, BAPOR and DONK, demonstrating whether or not they belong to the same crew.
2. A case study on the same graffiti artist who changes tag: MEAS and 12345. It is common for a graffiti artist to change their identity and sign with a different name. On many occasions this is done with the intention of avoiding criminal consequences. In the second case study, such a scenario is identified by using the presented colorimeter methodology.
3. A case study on the same graffiti artist who has used the same color to paint in different locations: two OVIS graffiti. The third case study presents the identification of authorship of the same author through the described technique.

After having carried out the first study presented in the first results subsection, several changes in the approach were proposed, taking into account the following aspects:

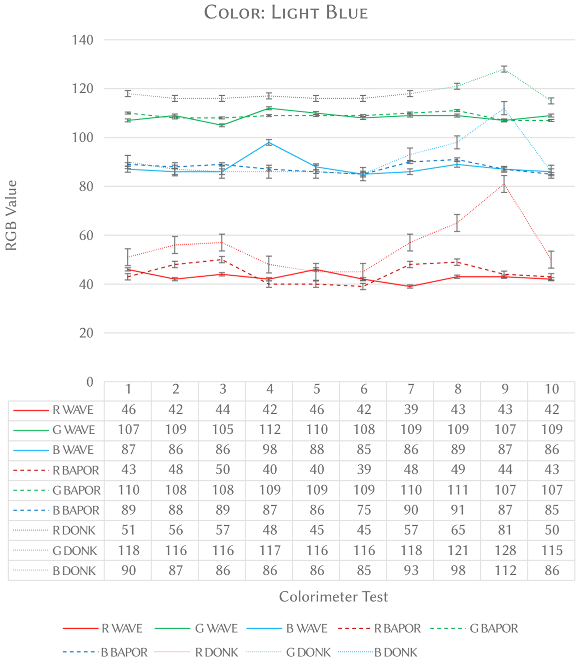
- Different locations around the city of Salamanca.
- In each case, 10 samples of the same color are taken with the colorimeter.
- The format of the obtained numerical data has been maintained, using statistical analysis.

A coincidence study has been carried out in which different colors from different graffiti have been selected. In those cases in which the colorimeter values coincided, it could be assumed that they had been made by the same crew or by the same author.

Case Study 1: WAVE, BAPOR and DONK Graffiti in the City of Salamanca.

The studies have been carried out selecting different colors and the result has been similar in all of them. In this case, the presented color was light blue and it was used to assess whether there was a colorimetric relationship between the three graffiti.

Fig. 19 shows the RGB values obtained by the colorimeter on the Y axis and the sample taken over time on the X axis. In addition, the integer values are shown in the table below and the error bars containing the standard deviation, indicating how spread out the data are around the sample mean. After comparing the values (RGB) obtained from the colorimeter we can conclude that there is a very close link between WAVE and BAPOR, but these same values are slightly further apart in the case of DONK. These results have been confronted with police sources and it has been confirmed that WAVE and BAPOR are from the same crew and not related to DONK.



	1	2	3	4	5	6	7	8	9	10
R WAVE	46	42	44	42	46	42	39	43	43	42
G WAVE	107	109	105	112	110	108	109	109	107	109
B WAVE	87	86	86	98	88	85	86	89	87	86
R BAPOR	43	48	50	40	40	39	48	49	44	43
G BAPOR	110	108	108	109	109	109	110	111	107	107
B BAPOR	89	88	89	87	86	75	90	91	87	85
R DONK	51	56	57	48	45	45	57	65	81	50
G DONK	118	116	116	117	116	116	118	121	128	115
B DONK	90	87	86	86	86	85	93	98	112	86

Fig. 19. Comparative table of RGB values taken as a sample, the "light blue" color in the graffiti WAVE, BAPOR and DONK.

Case Study 2: Graffiti "12345" Vs. Graffiti "MEAS" in the City of Salamanca

Case study 2 is a comparative experiment of two apparently different graffiti taking into account the descriptive color, which was different from the most common colors, namely, black, silver or white. In this case, the studied color is yellow to assess whether there is a colorimetric relationship between the two. Fig. 20 and Fig. 21 show the analyzed graffiti and Fig. 22, the values obtained in the study in the same way as in the first case.

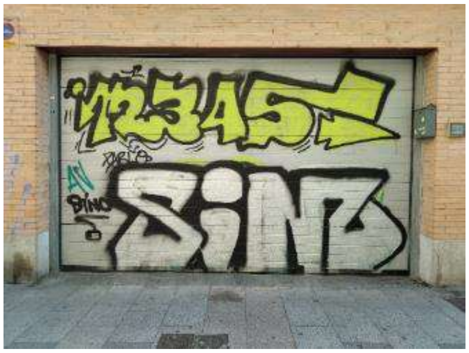


Fig. 20. Graffiti 12345.

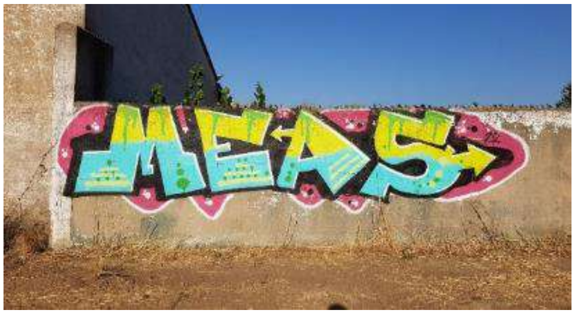
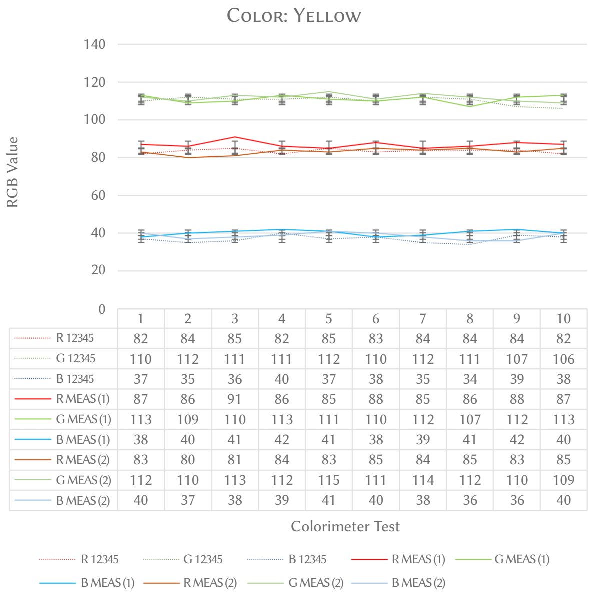


Fig. 21. Graffiti MEAS.

The premise to be taken into account is caused by police information that links the same person as the author of both styles (12345 would be the evolution of the MEAS tag) (12345 which is “12” like the “M”, “3” the “E”, “4” the “A” and “5 the “S”).



	1	2	3	4	5	6	7	8	9	10
R 12345	82	84	85	82	85	83	84	84	84	82
G 12345	110	112	111	111	112	110	112	111	107	106
B 12345	37	35	36	40	37	38	35	34	39	38
R MEAS (1)	87	86	91	86	85	88	85	86	88	87
G MEAS (1)	113	109	110	113	111	110	112	107	112	113
B MEAS (1)	38	40	41	42	41	38	39	41	42	40
R MEAS (2)	83	80	81	84	83	85	84	85	83	85
G MEAS (2)	112	110	113	112	115	111	114	112	110	109
B MEAS (2)	40	37	38	39	41	40	38	36	36	40

Fig. 22. Comparative table of RGB values taking as a sample the color “yellow” in graffiti 12345 and MEAS.

In this case, the initial premise is fulfilled by which the yellow color compared between both graffiti have the same colorimetric values. The conclusion according to the data obtained and the interpretation from this comparison is that it is the same author, or the same type of color has been used to paint with the same type of yellow in both cases.

Case Study 3: Two “OVIS” Graffiti in Two Different Locations in the City of Salamanca.

This comparative study shows two apparently equal graffiti considering that they share the same descriptive color and that it is different from the most common colors, namely, black, silver or white. In this case, the color being studied was light blue and the purpose of the study was to assess whether there is a colorimetric relationship between the two cases. Fig. 23 and Fig. 24 show the studied graffiti and Fig. 25 the results obtained. The premise in this case is to verify that the same author uses the same color in different pieces.

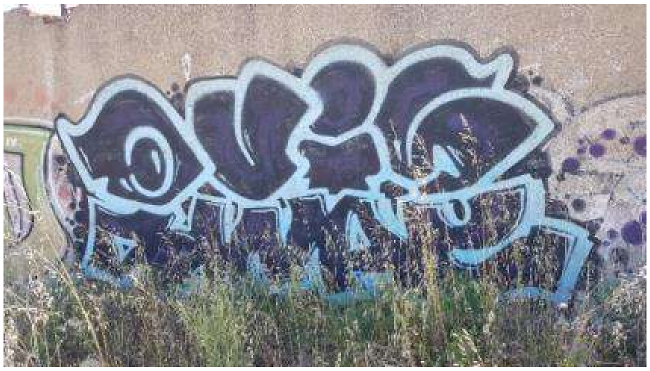


Fig. 23. Graffiti OVIS (1).

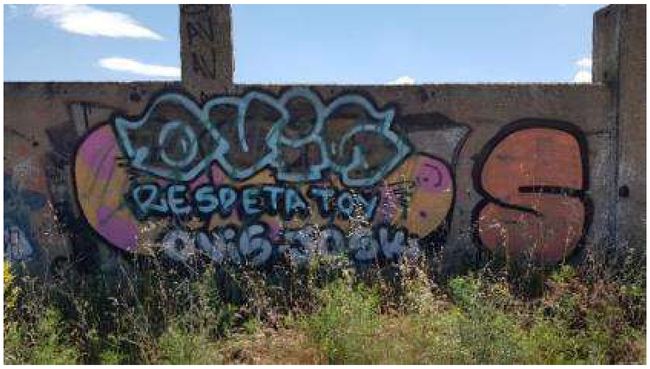
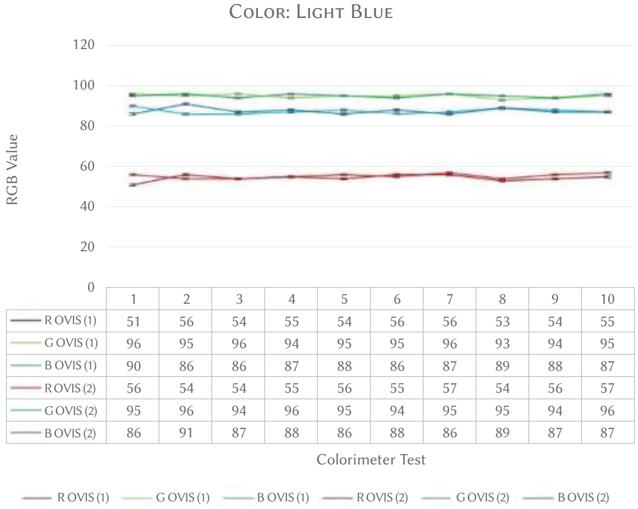


Fig. 24. Graffiti OVIS (2).



	1	2	3	4	5	6	7	8	9	10
R OVIS (1)	51	56	54	55	54	56	56	53	54	55
G OVIS (1)	96	95	96	94	95	95	96	93	94	95
B OVIS (1)	90	86	86	87	88	86	87	89	88	87
R OVIS (2)	56	54	54	55	56	55	57	54	56	57
G OVIS (2)	95	96	94	96	95	94	95	95	94	96
B OVIS (2)	86	91	87	88	86	88	86	89	87	87

Fig. 25. Comparative table of RGB values taking as a sample the color “light blue” in two OVIS graffiti.

In this case, the initial premise has been fulfilled by which the light blue color compared in these graffiti has the same colorimetric values.

According to the obtained data and the interpretation of this comparison, the conclusion is that it is the same author, or the same type of color has been used to paint the light blue color in both cases.

V. Conclusions

A. Functionality and Reliability of the Colorimeter

From the experiments carried out using the colorimeter, we obtain sufficient reliability to consider it as a suitable identification tool for this use (Fig. 26). To ensure a correct methodology, several samples were taken and compared with a color chart, which is a calibration method carried out prior to the experiments.

From these color shots it has been possible to observe that the colorimeter is a reliable tool. The colorimeter has provided certain “logical” values despite taking samples on a rough surface (wall), where the color saturation is different depending on the area and where the use of different materials could affect the glossiness of the surface. The obtained values have moreover been compared with the reference value of the color chart and small variations were considered normal as the measurements had not been taken in a controlled environment.

When it comes to capturing different shades of the same color, the colorimeter has performed correctly, as shown in the previous images. However, for lighter colors, such as sky blue, despite the fact that the device’s small bulb emits light of a similar color, when passing the numerical data from the colorimeter to an RGB reader, the obtained information was considered to be within the initial premise.

Colorimeter	Adobe Detection Tool	Colour Card
RGB: 48, 105, 89	RGB: 115, 217, 207	RGB: 143, 208, 189
🟩	🟩	🟩

Colorimeter	Adobe Detection Tool	Colour Card
RGB: 39, 110, 85	RGB: 85, 217, 204	RGB: 143, 208, 189
🟩	🟩	🟩

Colorimeter	Adobe Detection Tool	Colour Card
RGB: 30, 71, 136	RGB: 75, 98, 148	RGB: 65, 97, 144
🟦	🟦	🟦

Colorimeter	Adobe Detection Tool	Colour Card
RGB: 35, 74, 172	RGB: 30, 49, 108	RGB: 58, 95, 127
🟦	🟦	🟦

Fig. 26. Comparative table of the obtained color data.

Both the roughness of the wall, as well as the superimposition of some graffiti over others led to color differences in every centimeter of the graffiti as the wall had not been prepared in advance to carry out this type of graffiti. Even so, the colorimeter emits very similar results of the same color, regardless of the part of the wall in which we are and its state of conservation.

The use of the colorimeter has been validated as the data obtained for the same color on the same graffiti are within a range of stable values and this has made it possible to compare the same colors among different graffiti, with the aim of determining if they had the same colorimetric value.

B. The Colorimeter as a Tool For Identifying the Author of the Graffiti

As we have seen, through the colorimeter we find great similarities between the colors used in different graffiti. This leads us to raise the possibility that these graffiti have been done at the same time by a crew. Although we apparently think of different authors, it is possible to assume that the same author would have done both graffiti or that different authors (different graphical symbols) would use the same color spray to make their pieces independently, since in this case they are also arranged in different ways in a correlative manner.

In the first place, they could be two different authors -because that is what the graphical symbols point to however, they have used the same colors to create visual unity. Particularity that could be motivated by the fact that both belong to the same group. It should be noted that, without prior knowledge, as a result of the investigation, the calorimeter data led us to the conclusion that WAVE and BAPOR had similar image color values and experts confirmed that they were from the same Crew, which is one of the cases clarified by the police.

Secondly, the similarity in the use of color may be indicating that it is the same artist who has changed his style over time but has decided to keep his color palette. The contribution of this work is of great value in this sense when the same graffiti artist uses tags that do not share the same type as in the case study MEAS_12345. Graffiti involves a creation process in which drawings are made in sketchbooks or blackbooks [38] where the author practices the graphical symbols that they are going to use in their graffiti. It would be possible to think of the same author who, through this process, was capable of imitating or plagiarizing the signature of another author with the interest of "bombing" an area with graffiti of different authorship. Through the coincidences established in this study, we could think of it as a possibility for this case.

In conclusion, this work arises from the collaboration with a group of police experts in the field, who emphasize that the colorimeter study is a contribution to graphic analysis in the same way that the discrimination of inks followed in Graphistics can be. The analysis of inks used in graphistics is currently a highly valuable test in the courts and its aim is to verify whether two parts of the same document were written with the same pen or a different one, as well as whether two documents were written with the same writing tool. Therefore, the research presented here can provide, either by itself or by the combination of colors, an added feature to the graphistic analysis in the attribution of the authorship of a specific graffiti tag or the participation in a group action of the belonging crew.

C. Future Lines of Work

Based on the results of this research, future studies will compare different surfaces or supports, and different materials used in the elaboration of graffiti. A comparative study could also be carried out to contrast the performance of this proposal with other capture systems [39]. Other types of sensors could be applied, such as the spectrophotometer which can also be used in the industry to identify colors reliably through the amount of light absorbed by a sample, measuring the full color spectrum [40].

A future study that would further validate the premises set forth in this article would involve taking color samples and analyzing them chemically (gas spectro-photogrammetry) and thus obtain more precise data. Implemented within an identification, authorship, graffiti software would be a valuable step as it would be an additional variable to store in the database for pattern analysis using artificial intelligence [41].

Another possible investigation is the study of external values that influence image processing [42]. In this way, the data obtained with a colorimeter can be controlled and equated to that of a digital image.

Acknowledgments

This work has been supported by the project "XAI - XAI - Sistemas Inteligentes Auto Explicativos creados con Módulos de Mezcla de Expertos", ID SA082P20, financed by Junta Castilla y León, Consejería de Educación, and FEDER funds.

References

[1] F.J.A. Sanchís, "El postgraffiti, su escenario y sus raíces: graffiti, punk, skate y contrapublicidad", 2010, Doctoral Thesis.

[2] C. Castleman, "Getting Up/Hacerse Ver: El grafiti metropolitano en Nueva York", CAPITAN SWING S.L, 2012.

[3] H. B. L. Chi, D.N.N. Khanh, N.T.T. Vy, P.X. Hanh, T.N. Vu, H.T. Lam, N.T.K.P.L.Q. Hoang, "Development of a low-cost colorimeter and its application for determination of environmental pollutants", *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 249, 2021, doi: 10.1016/j.saa.2020.119212.

[4] A.W. Muhamada, A.A. Mohammed, "Review on recent Computer Vision Methods for Human Action Recognition", ADCAIJ: *Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 10, no. 4, pp. 361-379, 2022, doi: 10.14201/ADCAIJ2021104361379.

[5] B.P. Sharma and R.K. Purwar, "Ensemble Boosted Tree based Mammogram

image classification using Texture features and extracted smart features of Deep Neural Network", *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 10, no. 4, pp. 419-434, 2022, doi: 10.14201/ADCAIJ2021104419434.

[6] V.P. Mishra, "Texture Analysis using wavelet Transform", ADCAIJ*: Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 10, no. 1, pp. 5-13, 2020, doi: 10.14201/ADCAIJ2021101513.

[7] P.K. Patnaik, P. Mahapatra, D. Biswal, S.K. Nayak, S. Kumar, B. Champaty, K. Pal, "Development of a low-cost color sensor for biomedical applications", *Woodhead Publishing Series in Electronic and Optical Materials*, pp. 15-29, 2019, doi: 10.1016/B978-0-08-102420-1.00002-9.

[8] R. Valente, L. Barazzetti, "Methods for Ancient Wall Graffiti Documentation: Overview and Applications*", Journal of Archaeological Science: Reports.* Vol. 34, Part A, 2020, doi: 10.1016/j.jasrep.2020.102616.

[9] D. Wu, D.W. Sun, "Colour measurements by computer vision for food quality control - A review", *Trends in Food Science & Technology,* vol. 29, no. 1, pp. 5-20, 2013, doi: 10.1016/j.tifs.2012.08.004.

[10] G. Pellisa, M. Bertasaa, C. Riccib, A. Scarcellab, P. Croverib, T. Polia, D. Scalaronea, "A multi-analytical approach for precise identification of alkyd spray paints and for a better understanding of their ageing behaviour in graffiti and urban artworks", *Journal of Analytical and Applied Pyrolysis*, vol. 165, 2022, doi: 10.1016/j.jaap.2022.105576.

[11] H. Teng, A. Puli, M. Karakouzian, X. Xu, "Identification of Graffiti Countermeasures for Highway Facilities", *Procedia - Social and Behavioral Sciences,* vol. 43, pp. 681-691, 2012, doi: 10.1016/j.sbspro.2012.04.142.

[12] J.F. Ruiza, J. Pereirab, "The colours of rock art. Analysis of colour recording and communication systems in rock art research", *Journal of Archaeological Science,* vol. 50, pp. 338-349, 2014, doi: 10.1016/j.jas.2014.06.023.

[13] L. Llamas, "Medir valores RGB con Arduino y sensor de color TCS34725", 2018, enero, [Online]. https://www.luisllamas.es/arduino-sensor-color-rgb-tcs34725/

[14] CÓDIGOS DE COLOR HTML, 2021, [Online]. https://htmlcolorcodes.com/es/selector-de-color/

[15] ALLDATASHEET, Electronic Components Datasheet Search, 2022, [Online]. https://pdf1.alldatasheet.com/datasheet-pdf/view/894928/AMSCO/TCS34725.html

[16] Y.A. Badamasi, "The working principle of an Arduino", in 11 th International Conference on Electronics, *Computer and Computation (IECCO)*, 2014, doi: 10.1109/ICECCO.2014.6997578.

[17] J.C.S. Allende, J.H. Herranz, *"Una mirada al mundo arduino", Revista de Ciencia, tecnología y medio ambiente*, vol. 13, p. 28, 2015.

[18] L. Llamas, "Medir el color con Arduino y el colorímetro TCS3200", 2016, [Online]. https://www.luisllamas.es/medir-color-arduino-colorimetro-tcs3200/

[19] P.J.P. Fernández, "Diseño y construcción de un colorímetro de tres canales. Aplicación al estudio experimental de modelos de visión del color", 2004, Doctoral Thesis.

[20] P.D. Puiu, "Color Sensors and Their Applications" *Optical Nano- and Microsystems for Bioanalytics, Springer Series on Chemical Sensors and Biosensors*, vol. 10, p. 43, 2012, doi: 10.1007/978-3-642-25498-7_1.

[21] H. Torres, "HETPRO", 2017, [Online]. https://hetpro-store.com/TUTORIALES/sensor-de-color-tcs3200-con-arduino/

[22] BRICO GEEK, "Cómo construir un selector de color con Arduino", 2020, enero, [Online]. https://blog.bricogeek.com/noticias/arduino/como-construir-un-selector-de-color-con-arduino/

[23] AQ INSTRUMENTS, "Aquateknica", 2018, [Online]. https://www.aquateknica.com/que-es-un-colorimetro-como-funciona-y-para-que-sirve/

[24] A. Dahl, K. Jensen, J.M. Carstensen, K.C. Camilla H. Trinderup, "Comparison of a multispectral vision system and a colorimeter for the assessment of meat color," *Meat science*, vol. 102, p. 7, 2015, doi: 10.1016/j.meatsci.2014.11.012.

[25] V. Tomovic, I. Djekic, J. Miocinovic, B.G. Solowiej, J.M. Lorenzo, F. J. Barba, I. Tomasevic. B. Milovanovic, "Colour assessment of milk and milk products using computer vision system and colorimeter", *International Dairy Journal*, vol. 120, 2021, doi: 10.1016/j.idairyj.2021.105084.

[26] J. A. Coronado Martín, "El color, mediciones y aplicaciones.: Universidad Técnica del Norte", 2020.

[27] K. Minolta, "Sensing Konica Minolta", 2018, [Online]. https://sensing.konicaminolta.us/mx/blog/colorimetros-vs-espectrofotometros/

[28] R. Stojanovic, D. Karadaglic, "An optical sensing approach based on light emitting diodes", *Journal of Physics: Conference Series*, vol. 76, p. 7, 2007, doi: 10.1088/1742-6596/76/1/012054.

[29] Instructables Circuit, 2022, [Online]. https://www.instructables.com/Color-Sensor-1/

[30] Y. Shen, K. Yang, Ch. Lee, Ch. W. Neal, N. Xiong, "Color sensors and their applications based on real-time color image segmentation for cyber physical system", *EURASIP Journal on Image*, no. 23, p. 16, 2018.

[31] P. Luque, MJ. Fez, M.D. Capilla, "Colorímetro Colorlab: un colorímetro triestímulo para aplicaciones docentes", *Sociedad Española de Óptica*, Ed. Valencia, España, 2010.

[32] Ch. Ruiz López, T. Hoyer, A. Rebentisch, A.M. Roesch, K. Herkert, N. Huber, H. Floss, F. Juan, "Tool mark analyses for the identification of palaeolithic art and modern graffiti. The case of Grottes d'Agneux in Rully (Saône-et-Loire, France)" *Digital Applications in Archaeology and Cultural Heritage*, vol. 14, 2019, doi: 10.1016/j.daach.2019.e00107.

[33] I. Flores-Colen, J. de Brito, A. Dionisio, A. Moura, "Study of the cleaning effectiveness of limestone and lime-based mortar substrates protected with anti-graffiti products", *Journal of Cultural Heritage*, vol. 24, p. 14, 2017, doi: 10.1016/j.culher.2016.04.004.

[34] K. Kurata, "Open-source colorimeter assembled from laser-cut plates and plug-in circuits", *HardwareX*, vol. 9, p. 13, 2021, doi: 10.1016/j.ohx.2020.e00161.

[35] Instructables Circuits, 2022, [Online]. https://www.instructables.com/Arduino-Color-Detection/

[36] ADOBE COLOUR. 2021. [Online]. https://color.adobe.com/es/create/color-wheel

[37] X-RITE Pantone, 2022, [Online]. https://www.xrite.com/blog/colorimeter-vs-spectrophotometer

[38] F. Figueroa, "El "graffiti movement" en Vallecas historia, estética y sociología de una subcultura urbana (1980-1996)", 2003, Doctoral Thesis.

[39] S. Westland, "Imagen digital. Apuntes sobre diseño gráfico", 2001, [Online]. http://www.gusgsm.com/como_funciona_un_espectrofotometro_de_reflectancia

[40] AQ instruments. aqinstruments, 2015, [Online]. https://aqinstruments.wordpress.com/2015/10/08/colorimetros-y-espectrofotometros-comparacion-caracteristicas-funcionalidades-medicion-color/

[41] A.I. González, A.G. Arrieta, D. López-Sánchez. "Sistema inteligente en torno al mundo del graffiti", 2020, *Avances en Informática y Automática. Decimocuarto workshop.*

[42] V. Rajinikanth, S. Kadry, R. González-Crespo, E. Verdú, "A Study on RGB Image Multi-Thresholding using Kapur/Tsallis Entropy and Moth-Flame Algorithm", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 2, pp. 163-171, 2021, https://doi.org/10.9781/ijimai.2021.11.008.

Miguel García García

Miguel García García is Technical Engineer from the University of Salamanca. Later he began his specialization in the artistic discipline: he has a degree in Fine Arts (University of Salamanca), a Diploma in the Conservation and Restoration of Cultural Assets (specializing in Painting), a Graduate Degree in Conservation and Restoration of Cultural Assets (specializing in Graphic Documents), a Degree in Art History (National University of Distance Education), Master's Degree in Digital Animation (University of Salamanca), Master's Degree in Sacred Art (Pontifical University of Salamanca). He is currently a doctoral student, making it compatible with his business project (since 2015), restNOVA SOLUCIONES ARTÍSTICAS, dedicated to the Conservation and Restoration of Cultural Heritage.

Angélica González Arrieta

Angélica González Arrieta holds a PhD in Computer Science from the University of Salamanca and is College professor. She is also a member of the renowned research group BISITE (Bioinformatics, Intelligent Systems and Educational Technology), which was created in 2000. In 2021, the University of Salamanca awarded her the "Gloria Vegué" award for excellence in teaching. Her main line of research is the application of artificial intelligence in different fields: education, biometric identification, artificial vision, and data analysis. Regarding the publication and dissemination of results, the highlight is the set of 24 publications in international journals, of which 13 stand out with an impact factor, according to the JCR index (10 Q1, 3 Q2). Likewise, she has more than 50 publications in books and international congresses, some of them included in the CORE, SCI, etc. ranking. She collaborates on research issues with different companies in the private sector and also with public institutions such as the Ministry of the Interior. She has participated in multiple European (5), national (15) and regional (8) research projects. She has been the Principal Investigator of 12 teaching innovation projects. He has been awarded official recognition for two research periods of six years each.

Sara Rodríguez González

Sara Rodríguez is Associate Professor in the Department of Computer Science and Automation at the University of Salamanca. Her area of research is well-defined and focuses on the integration and use of multi-agent systems within the field of artificial intelligence; she formalized this line of research in her doctoral thesis in which an adaptive model for virtual organizations of agents was developed. Research activity can be considered the strong point of her curriculum, as shown by the 43 articles published in prestigious international journals and included in the Journal Citation Reports and her participation as a member of the scientific committee and speaker at international conferences. Her participation in numerous research projects over the years, including 5 international projects, demonstrates the transfer capacity and research quality endorsed by both companies and institutions. She has currently been awarded 2 six-year research periods and one six-year transfer period by the CNEAI. Her H index in Google Scholar is 39 with more than 4000 citations of her articles. Currently, she continues with her teaching and research activity within the GIR BISITE and as a member of the IoT Digital Innovation Hub of the USAL. Moreover, she continually progresses on the lines of Artificial Intelligence and Computer Science within the Department of Informatics and Automation of the USAL.

Sergio Márquez Sánchez

Sergio Márquez obtained the title of Industrial Engineer in 2017, from the University of Salamanca. At that same University he obtained his degree in Technical Engineering specializing in Mechanics (2014). In addition, he has a master’s degree in Finite Element Design, Drafting, and Analysis with Autodesk Inventor, SolidWorks, and CATIA V5. He has worked as a professor of robotics, programming and 3D design as well as participating in different entrepreneurship projects, R&D&i and product design in fields such as educational robotics and wearable technology. His research interests include smart textiles, electronic textiles, robotics, CNC technologies, PLM software, and circuit printing.

Carlos Fernando Da Silva Ramos

Carlos Ramos got his graduation from the University of Porto, Portugal, in 1986 and the PhD degree from the same university in 1993. He is Full Professor of the Department of Informatics at the Institute of Engineering – Polytechnic of Porto (ISEP-IPP). His main interests are Artificial Intelligence, Knowledge-based Systems, Decision Support Systems, Deep Learning, and Ambient Intelligence. He was the former Director of GECAD (Research Group on Intelligent Engineering and Computing for Advanced Development and Innovation), a research unit ranked with Excellent by the Portuguese Science and Technology Foundation, being currently member of the Administration Board. He is member of the Intelligent Systems Associate Lab from Portugal. He was Vice-President of the Polytechnic of Porto from 2010 to 2018, being responsible for the areas of R&D, Innovation and Internationalization, and since April 2022 he is Pro-President of the Polytechnic of Porto for Cooperation and International Relations. He has around 500 scientific publications (90 in scientific journals), participated in 58 international and national R&D projects, coordinating 25, and supervised 13 PhD completed works. Currently he is Director of the MSc Program on Artificial Intelligence Engineering from ISEP/IPP and member of the External Evaluation Commission of the Portuguese Agency for Evaluation and Accreditation of Higher Education for the area of Informatics for Polytechnics.

PeopleNet: A Novel People Counting Framework for Head-Mounted Moving Camera Videos

Ankit Tomar, Santosh Kumar, Bhasker Pant *

Graphic Era Deemed To Be University, Dehradun (India)

* Corresponding author. 87kumar.ankit@gmail.com (A. Tomar), amu.santosh@gmail.com (S. Kumar), pantbhasker@gmail.com (B. Pant).

Received 26 Mayo 2022 | Accepted 2 February 2023 | Early Access 12 April 2023

Abstract

Traditional crowd counting (optical flow or feature matching) techniques have been upgraded to deep learning (DL) models due to their lack of automatic feature extraction and low-precision outcomes. Most of these models were tested on surveillance scene crowd datasets captured by stationary shooting equipment. It is very challenging to perform people counting from the videos shot with a head-mounted moving camera; this is mainly due to mixing the temporal information of the moving crowd with the induced camera motion. This study proposed a transfer learning-based PeopleNet model to tackle this significant problem. For this, we have made some significant changes to the standard VGG16 model, by disabling top convolutional blocks and replacing its standard fully connected layers with some new fully connected and dense layers. The strong transfer learning capability of the VGG16 network yields in-depth insights of the PeopleNet into the good quality of density maps resulting in highly accurate crowd estimation. The performance of the proposed model has been tested over a self-generated image database prepared from moving camera video clips, as there is no public and benchmark dataset for this work. The proposed framework has given promising results on various crowd categories such as dense, sparse, average, etc. To ensure versatility, we have done self and cross-evaluation on various crowd counting models and datasets, which proves the importance of the PeopleNet model in adverse defense of society.

Keywords

Deep Learning, Density Map, Feature of Expansion, Moving Camera Videos, People Counting.

DOI: 10.9781/ijimai.2023.04.002

I. Introduction

Object detection and counting are emerging issues for the development of social sectors such as agriculture, wildlife sustainability, satellite imaging, drug molecule detection, crowd protection, etc. Computer vision (CV) facilitates pedestrian estimation to solve the social and administrative congestion monitoring problems which is a burning issue nowadays [1]. Automatic crowd counting is easier in video surveillance when the imager is stationary; however, it is more challenging in videos consisting of background motion driven by a moving camera and moving objects. Most crowd image databases are created by capturing from stationary cameras, resulting in a large number of occlusive samples, which limits the performance of any crowd counting (CC) mechanism. The existence of moving background, size of people, motion vibration, and camera position are some of the natural obstacles in crowd samples that also limit the people counting performance. There are some universally agreed challenges faced while developing an automated CC framework. Providing fair distribution of training information over live video streaming is the most common challenge [2] to protect the privacy of any individual by intentionally or unintentionally targeting individuals. Establishing a fair relationship between society and the monitoring system is another challenge to developing a reliable CC system. In addition, creating a simple and open-source crowd estimation model is also a hidden requirement for social welfare. With the mentioned challenges, moving camera surveillance proves to be very successful in places where static cameras are not installed due to any power, technical or geographical issues. Moving camera surveillance proves to be useful for our security to keep an eye on the enemy in difficult or high-altitude places.

Rapid urban population growth is a serious matter of concern to us, which is being worked upon by the research community nowadays. According to the world demographic report, 55% of the world's population today lives in urban areas, which will increase by about 70% in 2030 [3], as a result, our future will be surrounded by an unstructured and unbalanced crowd [4]. Stampedes, intentional gun firing, mob lynching, unnatural accidents, and unstructured traffic can have significant consequences for such disorderly population growth [5]. Moreover, the frightening worldwide casualties from 1975 to 2019[1] inspired us to develop an efficient crowd counting mechanism

[1] https://publications.iom.int/system/files/pdf/wmr_2020.pdf

Please cite this article as:
A. Tomar, S. Kumar, B. Pant, "PeopleNet: A Novel People Counting Framework for Head-Mounted Moving Camera Videos", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 6, pp. 61-73, 2024, http://dx.doi.org/10.9781/ijimai.2023.04.002

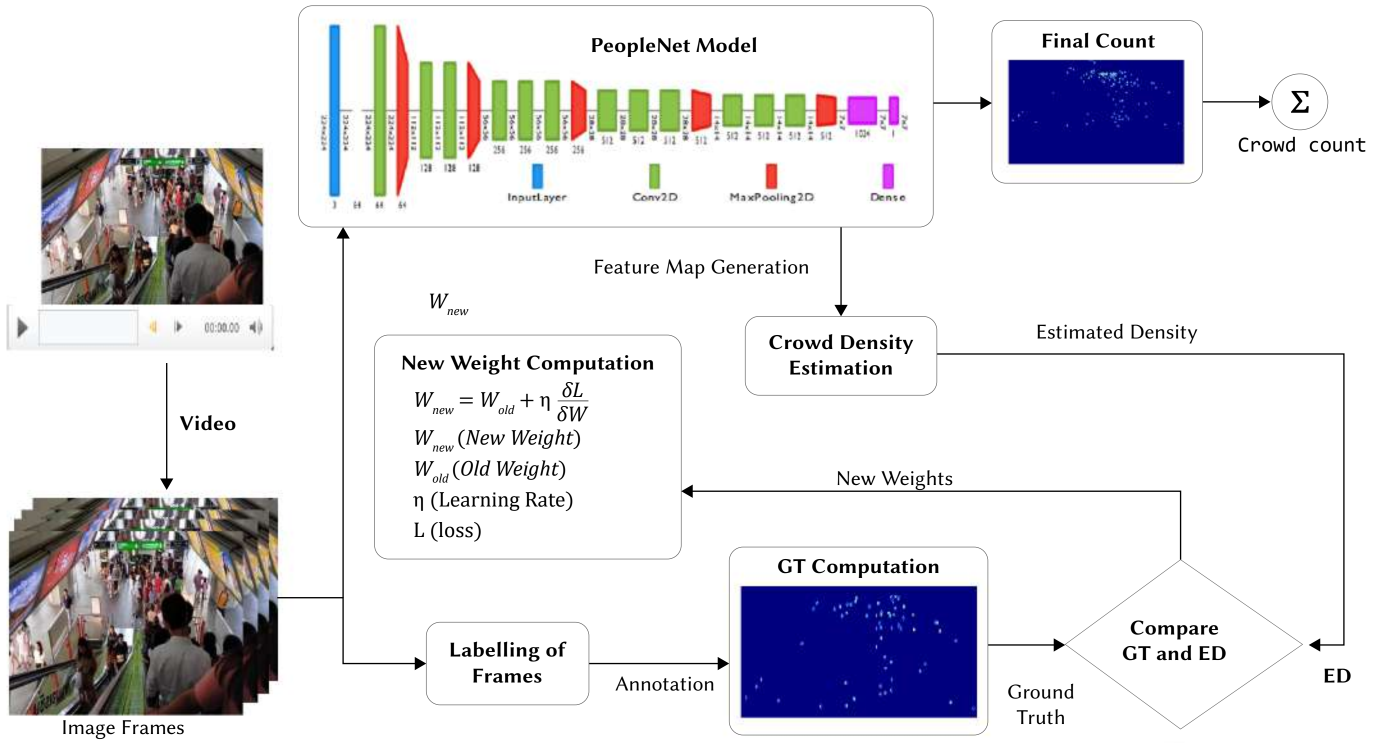


Fig. 1. Abstract working of People Counting framework.

in dynamic camera video surveillance environments. Modern CC approaches failed in accurate prediction of the crowd, resulting in lower accuracy; which also prompted us to work on advancing crowd counting via camera video surveillance.

Consecutive frame difference, context background subtraction, and optical flow are some traditional object detection techniques [6], [7], which often failed to handle some advanced pixel-level issues in motion parallax, moving background objects, blurring, and night vision. Optical flow is more robust among all but takes a longer time to produce real-time information than background subtraction [8]. Therefore, researchers are developing efficient and high-performing methodologies [9], [10] for automatic crowd counting nowadays via density map (DM) generation due to its wide spreading and exploring the video surveillance domain.

This study addressed the deep learning-based PeopleNet framework to handle the stated challenges for crowd counting in adaptive tracking camera environments. This framework follows a transfer learning CNN mechanism to perform pedestrian detection, which is designed by disabling the upper convolution blocks and replacing the last layers with new fully connected, and dense layers in the standard VGG16. Some significant feature extraction techniques are applied before inputting the crowded frames for generating good quality density maps (shown in Fig. 1). The construction of a crowd dataset in a moving camera environment and developing a lightweight, vision-based [11] pedestrian estimation framework is a significant contribution to this work.

The rest of the paper is organized into four sections. Related work is presented in section II. The proposed system along with detailed experimental work is discussed in section III. The obtained crowd counting results and state of art discussion are carried out IVth section, and finally, the paper is concluded in section V.

II. Related Work

Based on the performance, object detection techniques are divided into motion detection and motion estimation categories. The current literature work illustrates motion detection methodologies in traditional and advanced research levels, which is further described in traditional and advanced object counting methods.

A. Traditional Object Counting

Early object detection work accomplished by watershed segmentation algorithm [12], often suffers from feature spatiality and complexity issues, further solved the spatial segmentation by deploying color quantization via edge-preserving techniques [13]. Markov Random Field (MRF) model got better segmentation results [14], [15]; but limited to stationary video sequences. Jordan et al., [16] proposed an object tracking and detection scheme for both fixed and moving camera videos; however, it failed to solve the spatial ambiguity issue. Traditional object detection schemes are easy to implement but not intensive to advance challenges such as luminous variation [17], dynamic appearance [18], abrupt motion [19], occlusion [20], [21], complex background [22], [23], shadow [24] and camera motions [25] etc. The researchers embed transfer learning networks in traditional object detection techniques nowadays to handle these issues.

B. Advanced Object Counting

Deep Neural networks (DNNs) are more popular since 2016 for foreground detection [26], background subtraction [27], background generation [28] and deep spatial feature extraction [29], [30]. Guo and Qi [31] first developed Restricted Boltzmann Machines (RBMs) for moving object detection using background subtraction mechanisms. A deep auto-encoder was used by Xu et al. [32] for object detection in moving camera images; on the other side, a context encoder was presented by Qu et al. [33] for background subtraction as a backbone. Droogenbroeck [34], Cinelli [35] and Bautista et al [36] used CNNs for background subtraction.

Recent studies used two-stage [37], structured [38] and cascaded [39] CNN etc. Whereas a (NeREM) Neural Response Mixture and Mixture of Gaussian (MOG) [18] framework are employed to learn deep crowd features. Lempitsky et al. [10], first converted the labeled

```
Algorithm 1. Ground truth Generation
Input: A directory of .jpg files
T = 0, $I_C$;
Function ReadImage($readpath$, $(T+i)$, $I_C$);
        return readpath;
Function ImageResize ($l, b$);
        l ← HeightOfImage;
        b ← WidhtOfImage;
        return resizedImage;
Function WriteImage (($resizedImage$, [$readpath$,
$(T+i)$, $filepattern$)], $filepattern$);
        return writepath;
Function SaveMatFile ($writepath$);
        return mat;
while $i = 1$, $to\ n$ do
        call ReadImage();
        call ImageResize();
        call WriteImage();
        [x, y] ← getCoordinates;
        imageInformation.location ← [x, y];
        imageInformation.member ← [x, 1];
        call SaveMatFile( );
end
Output: A directory of .mat files
```

images into density maps (DMs) with the sum of a fixed Gaussian Kernel. Y Zhang et al. [40], proposed a fixed standard deviation to generate density maps. A single-column CNN [41] uses ResNet50 as a backbone for feature extraction. Boominathan et al., [42] introduced a patch-based MCNN network, which consists of each column with a different kernel size of the same depth as each parallel column. This model is able to produce high quality of density maps in pure deep learning environment; however fails to compute the outcome in efficient time. CSRNet [43] further overcame the stated problem via using two parallel shallow and deep networks to maintain the original density map resolution.

III. The Model Framework

High computation power requirement is one of the significant limitations of existing object detection mechanisms to tackle the dynamic background modeling implementation in real time. Lowering the computational overhead to get efficient crowd estimation results is a primary concern in the crowd counting techniques which had not performed efficiently for dynamic camera crowd videos. To tackle this issue of existing models, the PeopleNet model has proposed in this work. This model comprises the following components to estimate the pedestrian of the given video datasets.

1. Dataset Characteristics
2. Feature Generation
3. Feature extraction
4. The PeopleNet Architecture
5. Network Training
6. Crowd Counting
7. Performance Metrics

A. Dataset Characteristics

Promising work has been done in crowd counting form free and surveillance crowd samples [44]. Object detection problems can be solved in open CV, keeping static cameras into consideration; however not been practiced yet for dynamic objects. The unavailability of tracking camera video datasets is a significant need of the current research trend. So, we are obliged to construct an 1101 RGB image database prepared from the videos shot by a static or moving device. These videos have the different moving effects of the observing device and pedestrians. The dataset samples were shot in different places such as a mall, street, company corridor, restaurant, highway, escalator, roadside, tunnel, etc. in different timings, luminous appearances, shadows cast, in near-zero visibility that makes the dataset more realistic, practical, and overwhelming than existing ones. The crowd samples of videos of constructed datasets made public[2] for further research, of which the analogical description is shown in Table I.

TABLE I. Analogical Description of the Constructed Dataset

Dataset Attributes	Values
Resolution	1080x1920
Frame Rate (per second)	0.75
Total Samples	1101
Test/Train Samples	220:880
Crowd Variation	0-125
Total Pedestrians	33567
Color/Format	RGB/JPG
Place	Multiple Locations
Property	Walking, Eating
Average Crowd Size	31
Shadow/Reflection/Loitering	Yes

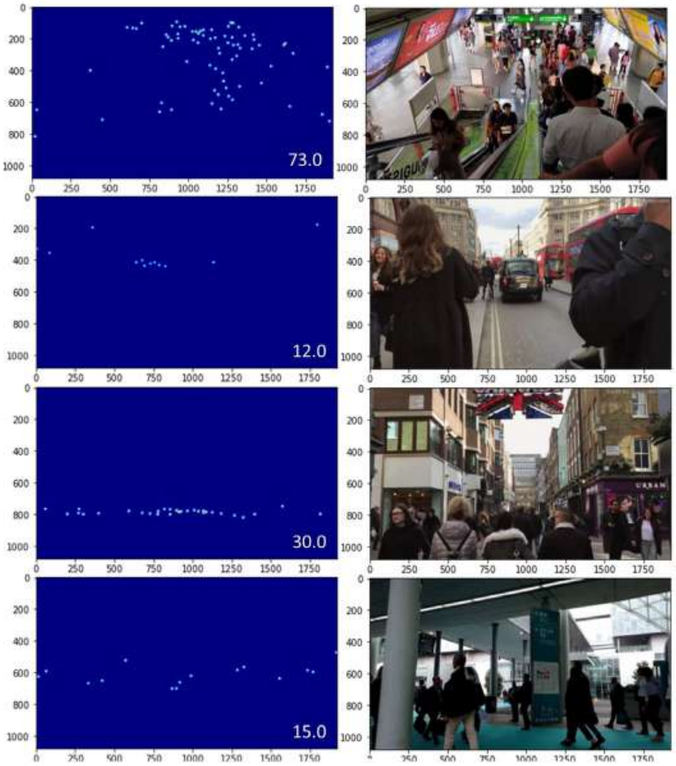


Fig. 2. Crowd samples with their respective density maps.

[2] https://www.kaggle.com/ankit87/moving-camera-dataset

B. Feature Generation

The feature generation is an indivisible part of data preprocessing before input the information. The crowd features are generated via ground truth and density map generation, its whole process has discussed through pictorial and algorithmic way.

Ground Truth and Density Map Generation: Ground truth generation from the crowded images is based on current research trends [45], [46]. This work formulates the people estimation via density map via density regression function. The image frames and respective DMs (obtained from crowd video clips) are the training input of PeopleNet model. A DM of object detection is obtained from x, and y coordinates of the people's head location also called GT labels. The original image (bottom right portion), annotated samples (Left portion), and DM (Upper Right portion) are shown in Fig. 3. Consider the $I_i = (I_1, I_2, \quad , I_{TS_n})$ are the image frames obtained (training samples) from crowded videos. The ground truth label $n^{GT,i} = (n_1, n_2, ..., n_{X_i})$ for center point X represents each people's head presented in crowded samples; which is obtained via density map $D^{GT,i}$, described through (1).

$$\forall p \varepsilon I_i, D^{G_T,i}(p) = \sum_{p \varepsilon n_i^{G_T}} G\, D^{G_T}(p; \mu = n^{G_T,i}, \sigma^2) \tag{1}$$

The GD^{GT} is Gaussian distribution (σ) for 'p' pixels. The total crowd count in sample Ii is obtained by summing the density values for all pixels described in (2) and samples are shown in Fig. 2.

$$G_T = \sum_{p \varepsilon I_i} D^{G_T,i}\,(p) \tag{2}$$

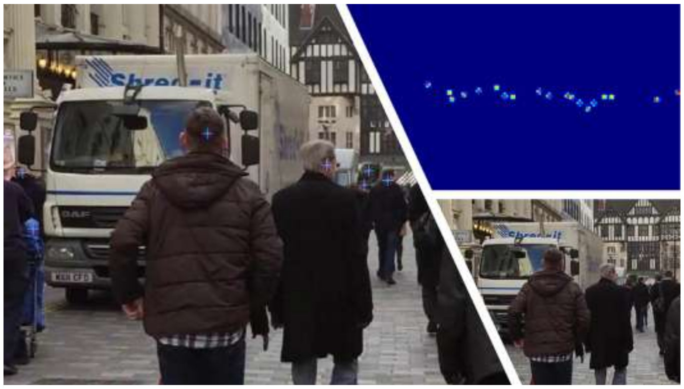


Fig. 3. Obtained Image samples from Video (Right Lower), Annotation Creation (Left), and, generated Density Map(Right Upper) from the image sample.

C. Feature Extraction

To tackle the poor crowd counting performance due to high computation power demand, this work incorporated the focus of expansion (FOE) concept in feature preprocessing. The FOE plays a significant role in accurate flow estimation for CV applications such as range & obstacle estimation. The field effects of FOE signify the transformation and rotation motion caused by the dynamic camera. Efficient crowd feature (segment, edge-based, and texture) selection incorporating the feature of the expansion concept via diverging optical flow vectors to estimate the motion fields depicted in Fig. 4.

Consider the $\vec{V} = (R_x, R_y, R_z)^T$ camera motion towards $P = (x, y, z)^T$ fixed points, where FOE is computed (x_{FOE}, y_{FOE}) against the pixel corresponding to the P crowd image plane. At a particular point of an image, FOE is obtained by the intersection of the image plane and camera motion velocity, while the camera is in relocatable motion [47].

$$V_x = \frac{R_z x - R_x f}{Z}, V_y = \frac{R_z y - R_y f}{Z} \tag{3}$$

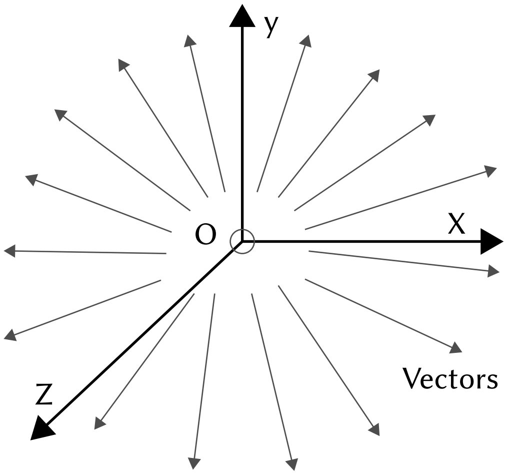


Fig. 4. FOE Diverging Optical flow vectors in temporal crowd.

A 3D coordinate system defines X, Y, and Z planes with the optical axis of the camera, which is parallel to the Z-axis and, X, Y-axis is parallel to the image plane for a particular location L(X, Y, Z) in a 2D plane at projection P(x, y, z) for the 3D plane [48]. The velocity V is derived under 3D space defined in (3), where V_x, V_y are velocity plane vectors. R_x, R_y, R_z are relocatable 3D components for focal distance f. Defining FOE in (4)

$$x_0 = \frac{fR_x}{R_z}, y_0 = \frac{fR_y}{R_z} \tag{4}$$

$$V_x = (x - x_0)\frac{R_z}{Z}, V_y = (y - y_0)\frac{R_z}{Z} \tag{5}$$

The (3) becomes in a linear system with (x_0, y_0) focus of expansion in (5) further used in (6).

$$\begin{bmatrix} V_{Y_1} & V_{Y_n} \\ V_{X_1} & V_{X_n} \end{bmatrix} * [x_0, y_0] = \begin{bmatrix} V_{Y_1}V_{X_1} & V_{Y_n}V_{X_n} \\ V_{Y_k}V_{X_k} & V_{Y_{nk}}V_{X_{nk}} \end{bmatrix} \tag{6}$$

The FOE detection is based on the above properties including flow and matched filter with size (2w+1) × (2w+1), in Fig 4, having each pixel shows the angle between the origin and grid point (7).

$$f(x, y) = archtan\left(\frac{x}{y}\right) - w \le x \le w, -w \le x \le w \tag{7}$$

For the given images $I_1\,(x, y)$ and $I_2\,(x, y)$, $\Delta t \rightarrow 0$ time apart, assume FOE based optical flow can be obtained corresponding to the flow of x, y-axis. Furthermore, the optical flow is tuned with segmented, edge-based, and texture features.

- **Segmented Features**: The segmented features capture foreground entities (blob, shape and size) at reference pixels for density map $D^{GT,i}$, of mathematical expression is described in (8).

$$S = \sum_{n=0}^{p} S_n, where S_n = \sum_{(x,y) \in P_n} \sqrt{D^{G_T,i}(x, y)} \tag{8}$$

- **Edge-Oriented Features**: Consist Minkowski dimensions to estimate strong crowd counting ability via (9).

$$e = \sum_{n=0}^{p} e_n, where\ \ e_n = \sum_{(x,y) \in P_n} \sqrt{e(x, y)} \tag{9}$$

- **Local-Texture Features**: These features are employed for density classification across the crowded regions r, depicted in (10).

$$g(r) = \sum_{(x,y) \in r_n} 1, for Q_z(x, y) = Q_i(x', y') \tag{10}$$

D. The PeopleNet Architecture

In deep learning, CNN is enough capable of automatic feature extraction and prediction process, which evolved into transfer learning. The PeopleNet model is capable of crowd estimation using five convolution groups of 21 layers, initially, image samples were provided in batches with three RGB channels, as shown in Table [tabsecond]. The baseline network of this work is obtained from VGG16 architecture by stacking max pooling, convolution, dense, and fully connected (FC) layers. Two significant changes have been made to the standard VGG16 network to make it fit for people counting purposes, first, we have disabled its top seven layers by freezing them (making their status 'false') and added our dense layers by replacing its FC layers. The input videos have an original resolution of 1080×1920, which further took 224×224 after preprocessing and normalizing. The input stream passed through Conv2D groups described by the following changes:

1. For the 1st Conv2D group, double convolution layers with 64 filters of [3×3] with a stride of (1, 1) have been applied followed by a 2D pooling layer of size [2×2] with stride (2, 2). However, we have frozen the first convolution group by setting their status '*False*'.
2. In 2nd Conv2D group, a double layer with 128 filters of [3×3] and stride (1, 1) have convolved over the output of the previous layers. We obtained a 2D pooling of 128 kernels after applying a stride of size (2,2).
3. The 3rd Conv2D group is composed of three layers having 256 filters of [3×3] dimensions with the stride of (1, 1). A stride (2, 2) and a max-pool [2×2] layer have been used in the sub-sampling technique.
4. The 4th Conv2D group has three consecutive convolution layers [3×3] with 512 filters in each and a stride (1,1) with pooling layers of size [2×2].
5. Likewise, the last and 5th group has convolution layers of 512 filters that have been convolved three times with the stride of (1, 1).
6. Finally, two fully connected layers of size 1024 and 1 are deployed at the end.

The proposed model works with n image frames I of M_i dimension matrix. A kernel K matrix is convolved through each image to create feature maps through equations (11) and (12).

$$s(i,j) = \sum_{m}\sum_{n} I\,(m,n)K(i-m)(j-n) \tag{11}$$

where

$$I[m,n] = \sum_{m}^{s}\sum_{n}^{t}(X[m+s][n+t]).C[s][t] \tag{12}$$

Where y is the output image, I image frame, C convolution mask, and t tokens. The value of K is taken 3 to carry out this experiment for $W \times H$ image dimensions for the p pooling matrix. The Euclidean loss is replaced by average pooling in (13) to estimate the N_i^{GT} as ground truth for spatial units U.

$$N_i^{GT} = \frac{1}{U}\sum_{j}\hat{y}_i\,(x_j) \tag{13}$$

A customized loss function is required to train the effective DL models, which is derived from the difference between actual and estimated count (EC), depicted in (14).

$$W_{new} = W_{old} - \eta\frac{dL}{dW} \tag{14}$$

Where (x_i, y_i) are spatial coordinates for density map. W_{new} and W_{old} are the updated and older neuron weights gained at each forward and back prorogation with the help of learning rate η. The actual and estimated people counts are used to compute the pixel-level Euclidean distance loss function $L_D(\theta)$, which is defined in (15).

$$L_D(\theta) = \frac{1}{N_{ts}}\sum_{i=1}^{N_{ts}}(\frac{H_C)(I_i;\theta) - G_T}{G_T+1})^2 \tag{15}$$

A set of learn-able metrics (θ) derived from total sample parameters are N_{ts}. The actual head-count is denoted by G_T, and the estimated head-count is denoted by $H_C(I_i; \theta)$. The PeopleNet is trained from the scratch for random network parameters and resolves poor performance issues for a sparse crowd using $L_D(\theta)$, which helps to meet the real-time computation. The loss in (16) is given by

$$\underset{\theta}{\nabla} L_D = \{0; if N_i^{G_T}(I_i) - N_i^{G_T}(I_{i-1}) + \varepsilon \leq 0 \underset{\theta}{\nabla} N_i^{G_T}(I_i) - N_i^{G_T}(I_{i-1}) \tag{16}$$

A popular mean of square error (MSE) loss is used to train the model via computing it between estimated and real pedestrian values. The experiment conduction used *Adam* as an optimizer function since it is best suitable for non-moving objects for noisy/sparse gradients, also provides a regret bound on convergence rate comparable to the convex optimizer eloberated in below equation (17).

$$\Theta_{t+1} = \Theta_t - \frac{\eta}{\sqrt{\widetilde{v}_t + \epsilon}}\hat{m}_t \tag{17}$$

Here $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$ and $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$, are initialized vectors and biased towards 0, where β_1, β_2 are close to 1, whose default values are taken 0.9 and 0.999 respectively. The density map learning paradigm D(p) uses a set of X and Y parameters via mapping function F using θ as a parameterization function to predict the labels in (18).

$$D(p) = \int (F(X,\theta) - N^2)p(X,N)p(X,Y) \tag{18}$$

After each layer, the output is defined in (19); where O_i is output f is filtering, P is padding unit, and S is stride.

$$O = \frac{O_i - f + 2P}{S} \tag{19}$$

The proposed methodology is unique and uses shortcuts to all preceding blocks from each convolutional group. As a result, a combination of local and global features is fed to the first and last layers. We included padding, stride averaging, and regularization techniques for feature map matching.

E. Network Training

CNN-based people counting frameworks have many trainable parameters or complex architectures, which makes model training more difficult and the feeding process time-consuming. In this study, the PeopleNet model is performing better on fewer trainable parameters in less computation. For the given training set of 880 images and their respective DMs, two counter variables F and G train the convolution model jointly with the help of L_D (θ). Algorithm 2 summarizes the training procedure for the proposed model. Let (I_i, D_i) be the pair of i^th image and density maps, and (F_{I_i}) (G_{D_i}) are original and refined density maps respectively, the loss (L) can be expressed through (20).

$$L(\theta)_D = \sum_{i=1}^{N_{ts}} |F(I_i) - G(D_i)|^2 + \alpha|G(D_i) - D_i|^2 \tag{20}$$

$|G_{D_i} - D_i|$ helps in training and predicting the refined density maps during every back and forward prorogation, and the loss tends to converge to the minimum level for smooth training. The training accuracy and validation loss statistics for 300 epochs of PeopleNet are shown in Fig. 5.

TABLE II. Layered Architecture of PeopleNet Framework.

Layer Type	Output Shape	Parameters	Trainable Status
Conv2D	64 × 64 × 224 × 224	1792	False
BatchNorm2D		128	False
Relu			
Conv2D		36928	False
BatchNorm2D		128	False
Relu			
MaxPoll2D			
Conv2D	64 × 128 × 112 × 11	73856	True
BatchNorm2D		256	True
Relu			
Conv2D		147584	True
BatchNorm2D		256	True
Relu			
MaxPoll2D			
Conv2D	64 × 256 × 56 × 56	295168	True
BatchNorm2D		512	True
Relu			
Conv2D		590080	True
BatchNorm2D		512	True
Relu			
Conv2D		590080	True
BatchNorm2D		512	True
Relu			
Conv2D	64 × 512 × 28	1110860	True
BatchNorm2D		1024	True
Relu			
Conv2D		2359808	True
BatchNorm2D		1024	True
relu			
Conv2D		2359808	True
BatchNorm2D		1024	True
Relu			
MaxPooling2D			
Conv2D		2359808	True
BatchNorm2D		1024	True
Relu			
Conv2D		2359808	True
BatchNorm2D		1024	True
Relu			
Conv2D		2359808	True
BatchNorm2D		1024	True
Relu			
MaxPooling2D			
AdaptiveAveragePooling2D			
Flatten			
BatchNorm2D		2048	True
DropOut			
Linear	64 × 512	524288	True
Relu			
BatchNorm1D		1024	True
DropOut			
Linear	64 × 1	512	True

F. Crowd Counting

The position of each person's head is labeled with white cross symbols (H_x, H_y) as a delta function (Fig. 3); which is used to compute the labeled images into a DM by convolving operation through G_{α_i}. The inverse KNN distance method obtains labeled values from G_T by computing KNN distance from p(X_x, X_y) to $p(X_x, X_y)$ pixel values, for H people heads (depicted in algorithm 2). The data samples and respective annotated DMs are parallel inputs provided to the PeopleNet model, which can generate final maps after 8 hours of intensive training.

```
Algorithm 2. PeopleNet Training Procedure
Input: A pair of Images and respective density maps
        $(I_i, D_i)_{i=1}^{N}$
Initialize two F and G counters;
for Epoch ← 1...... $N_E$ do
    for Epoch ← 1..., .. N do
        Estimate DM ($F_{I_i}$).
        Generate $G_T$ ($G_{D_i}$).
        Update F counter using loss L in (20)
        update $N_G$ for every epoch through counter G
        if mode(Epoch, $N_G$) == 0 then
            update the parameter using $L_D(\theta)$, (15)
        end
        else
        end
    end
end
Output: Updates values of counters F and G.
```

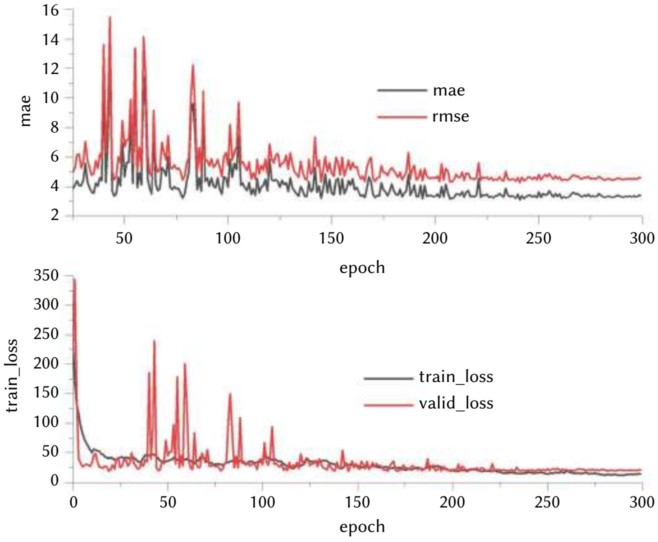


Fig. 5. Training accuracy and loss statistics of PeopleNet.

In the experimental procedures, it has been observed that the obtained result is more precise upon disabling the first convolution layers than the other layers for the stated loss function (Fig. 6). In this figure, the first two rows have the original image and original DMs, the next rows show the output generated DMs (printed with sequence, PSNR, G_T, and E_C).

G. Performance Metrics

After widely investigating the universally accepted articles, we have divided the performance evaluation into the image and pixel-level categories [24]. The quality of generated DMs is used to evaluate the pixel-level performance, whereas popular regression model enumerate the image-level performance.

1. Image Level Error

Root Mean Squared Error (RMSE) and, Mean absolute error (MAE) are commonly used as image-level accuracy measurement metrics [49], which compute the overall deviation between the estimated and actual samples values. The mathematical expressions of RMSE and MAE can be seen in (21) and (22).

$$RMSE = \sqrt{\frac{1}{N_{ts}}\sum_{i=1}^{N_{ts}}(C_i^{GT} - C_i^{Ec})^2} \tag{21}$$

$$MAE = \frac{1}{N_{ts}}\sum_{i=1}^{N_{ts}}(C_i^{GT} - C_i^{Ec}) \tag{22}$$

RMSE is a more corrective measure to assess the insignificant sample deviation; however, MAE generally fails to secure the overall accuracy for huge variation data samples. Therefore, mean absolute percentage error (23) would be the better measurement choice.

$$MAPE = \frac{1}{N_{ts}}\sum_{i=1}^{N_{ts}}\left|\frac{C_i^{GT} - C_i^{Ec}}{C_i^{GT}}\right| \tag{23}$$

MAE, MAPE, and RMSE define robustness and global image-level accuracy. To evaluate local region accuracy, correlation coefficient r is another better alternative to measure the dependence between two variables (24), where $A = C_i^{GT}$ and $B = C_i^{Ec}$.

$$r = \frac{N_{ts}(\sum(A)(B)) - (\sum(A))(\sum(B))}{\sqrt{[N_{ts}\sum(A)^2 - (\sum(A))^2] * [N_{ts}\sum(B)^2 - (\sum(B)^2]}} \tag{24}$$

To measure the covariance, association or statistical relationship of two continuous variables, the Pearson correlation coefficient (P_r) is another significant test statistics, which is the fundamental measure the strength of the linear relationship shown in (25). Sometimes it is also called as coefficient of determination.

$$P_r = \frac{\sum_{i=1}^{N_{ts}}(C_i^{GT} - \overline{C_i^{GT}})(C_i^{Ec} - \overline{C_i^{Ec}})}{(\sqrt{\sum_{i=1}^{N_{ts}}(C_i^{GT} - \overline{C_i^{GT}})^2}\sqrt{\sum_{i=1}^{n}(C_i^{Ec} - \overline{C_i^{Ec}})^2}} \tag{25}$$

Adjust R_A^2: R^2 often suffers from score improvement in increasing terms, even if the model improvement remains the same, which might create a misguidance for the researchers. Therefore, R_A^2 is used to improve in case of any real improvement via adjusting the increased estimators, in (26) k is independent variables for n observations in operation.

$$R_A^2 = 1 - (1 - R^2)\left[\frac{N-1}{n-1-k}\right] \tag{26}$$

The Normalized Root Mean Square Error (NRMSE) (27) use to facilitates the comparison between models with different scales to interpret as a fraction of the overall range that is typically resolved by the model.

$$nrmse = \frac{RMSE}{\overline{C_i^{Ec}}} \tag{27}$$

$\overline{C_i^{Ec}}$ is the average of observation value computed (21).

2. Pixel Level Error

PSNR (Peak signal-to-noise ratio) [50] is the most common pixel-level metric used to measure the error deviation between

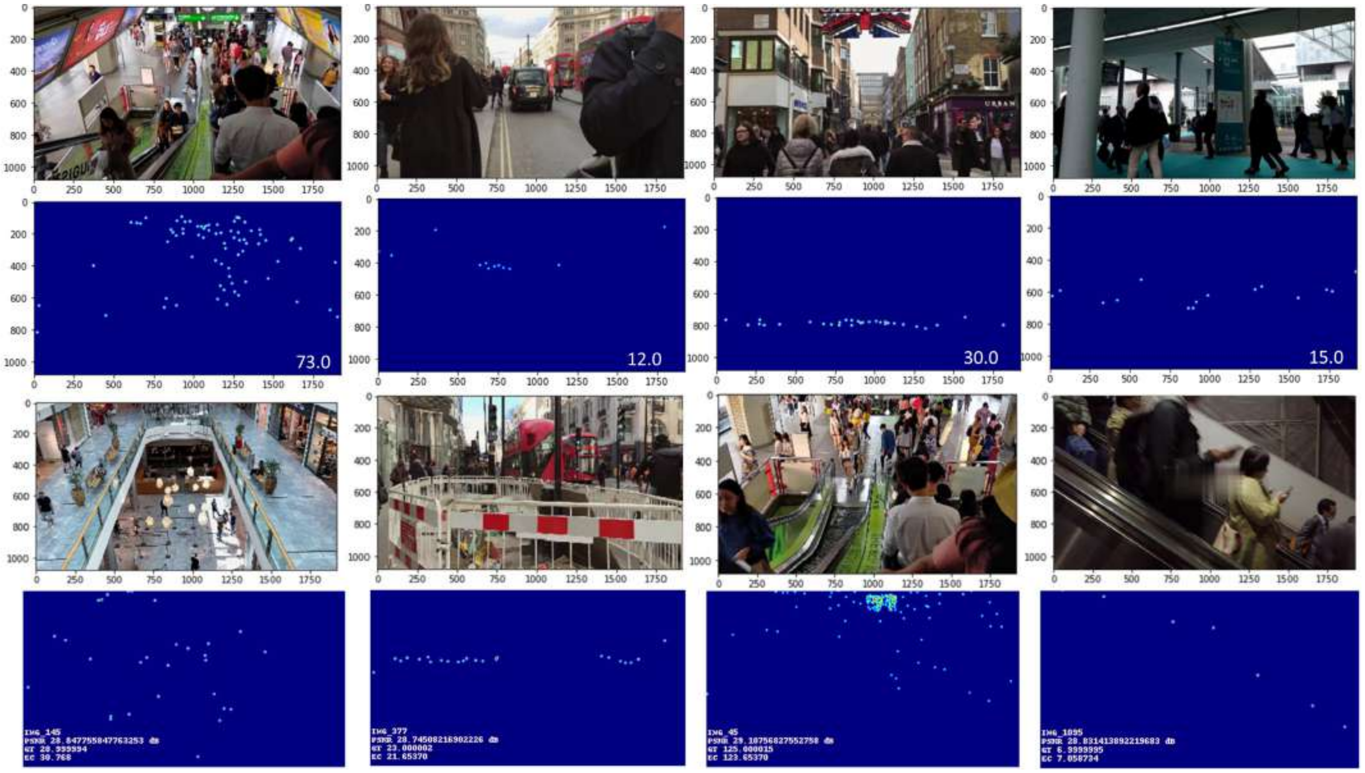


Fig. 6. Frames and their respective density maps are shown in first two rows, while last two rows represent testing output frames and generated density maps printed with some information viz image-sequence, PSNR value, GT and EC, etc.).

TABLE III. Performance Analysis of Proposed Model Different Category Samples

Category	Total Frames	Crowd Size	MAE	RMSE	Pr	r	nrmse	SSIM
Dense	165	51-125	15.514	19.595	0.890	0.944	0.124	0.31
Sparse	260	0-15	12.24	15.453	0.912	0.932	0.131	0.37
Average	676	16-50	9.376	12.72	0.907	0.933	0.156	0.35
Overall	1101	0-125	3.43	4.623	0.917	0.919	0.166	0.34

corresponding and original DM pixels. The resolution size of both the original and degraded image matrix must be the same while working with the 2D matrix.

$$PSNR = 10log_{10}(\frac{MAX_f}{\sqrt{MSE}}) \tag{28}$$

In (28), 10log defines the square of amplitudes in terms of noise and, MSE is defined in (29).

$$MSE = \frac{1}{M * N}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} | O(i,j) - D(i,j)|^2 \tag{29}$$

Where O and D are data matrices of original and degraded DMs. MAX_f denote maximum signal value for M rows denoted by i and N columns represented by j pixels. Structural Similarity Index: SSIM([50]) is used to measure the perceptual difference between two identical images this parameter cannot judge the better of two images.

IV. Result and Discussion

This section mainly focuses on the validation efficacy of PeopleNet architecture for a variety of crowd scene categories. The training and accuracy error deviation is expressed separately through Fig. 5 and Fig. 7. Moreover, the comparison of the cross-scene results is comprehended separately for existing state of art CC models and benchmark datasets.

A. PeopleNet Model Performance Analysis & Validation

The crowd estimation results are validated over the test data samples, whose predicted results are computed and shown graphically. A thorough, deep, and strategic error deviation of this experiment has been done in three main parts as follows:

1. Scenario Error Estimation

Table III presents the results obtained from the PeopleNet framework on the constructed dataset. The data samples have been categorized into dense, sparse, and average scenarios. The framework is tested separately by modeling in each category to better analyze the people counting error. The dense category data subset contains 165 frames having a people range from 51 to 125, over which the PeopleNet model secures the MAE, RMSE, Pr , r, nrmse and SSIM as 15.514, 19.595, 0.890, 0.944, 0.124 and, 0.31 respectively. The sparse crowd category kept a total of 260 samples (having people ranging between 0 to 15) and the model secured 12.24, 15.45, 0.912, 0.932, 0.131 and 0.37 as MAE, RMSE, *Pr*, *r*, nrmse and SSIM respectively. Moreover, there are 676 average category samples of 16-50 crowd size and secured 9.376, 12.72 as MAE and RMSE.

Satisfactory performance is observed for sparse category samples; insufficient training samples are the main cause of lower accuracy. Instead of training over enough samples, the model achieved satisfactory results for average crowd samples by securing average MAE and RMSE. This accuracy is low as compared to the overall category dataset. The proposed model performance is superior in the overall category dataset sample with 1101 frames of people ranging from 0-125 and achieved MAE 3.43 and RMSE 4.623 respectively with SSIM 0.34. A slight difference is observed in model performance from various crowd categories, which indicates that the proposed model neither performs poorly for dense nor outperforms for sparse samples. The main aim to present the Table III is to test the PeopleNet's robustness for low and high-crowded samples.

2. Training Error Estimation

Existing CC research focused on the accuracy parameters only, but the proof of how well the model training performed is still missing in the literature. Here we exposed the training performance through pictorial and factual representation. During training, the difference between MAE and RMSE is higher in the first 50 epochs; however, it is less afterward but fluctuates with a very high margin. The fluctuation lowered between 100 to 200 epochs. After 200 epochs, a negligible fluctuation is observed as both MAE and RMSE are parallel to one another from 150 epochs till the end. These statistics ensure a smooth training process without over, or under-fitting.

3. Crowd Error Estimation

The performance of any AI model mainly depends on its architecture, hyperparameter setting, quality, and quantity of training samples. In this experiment, the obtained results vary from test-to-test samples as they may belong to a different video clip. The accuracy deviation between estimated and real people obtained by the proposed model over 220 different image frames could be seen in Table IV.

TABLE IV. Summary of Accuracy Deviation Obtained by the Proposed Model Over 220 Image Frames

Frame Number	Frame Image	Ground Truth	Estimated Count	Deviation Accuracy
1^{th}	f_1	29	26.383	2.617
2^{nd}	f_2	29	31.4189	-2.4189
——	——	——	——	——
10^{th}	f_{10}	39	49.2062	-10.2062
11^{th}	f_{11}	06	7.3051	-1.3051
——	——	——	——	——
50^{th}	f_{50}	14	12.0589	1.9411
51^{th}	f_{51}	27	27.4985	-0.4985
——	——	——	——	——
100^{th}	f_{100}	69	64.2111	4.7889
101^{th}	f_{101}	23	27.5823	-4.5823
——	——	——	——	——
150^{th}	f_{150}	101	86.4342	14.5658
151^{th}	f_{151}	04	6.0725	-2.0725
——	——	——	——	——
200^{th}	f_{200}	31	28.1497	2.8503
201^{th}	f_{201}	23	22.8873	0.1127
——	——	——	——	——
119^{th}	f_{119}	03	05.1674	-2.1674
220^{th}	f_{220}	41	48.7397	-7.7397

Furthermore, the same results are also depicted through Fig. 7, where a yellow slider area is mapped to render the zoomed view. For each edge, the quantified results have been associated, the absolute value represents C_i^{GT} and the fractional values represent C_i^{Ec}. The underestimation CC effect can be observed for more than 100 people. Insufficient training samples for dense crowd samples may be the leading cause; however, the nearness of obtained results and parallelism of lines is evidence of the model's outstanding results.

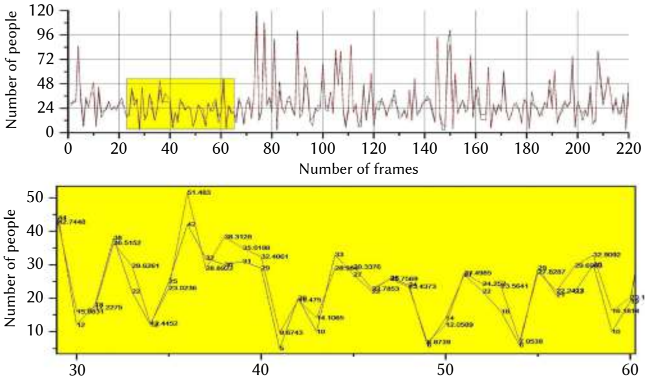


Fig. 7. Accuracy deviation illustration between estimated and actual crowd count.

B. Comparative Analysis

This work focuses on the result variation, model behavior, and data variance on moving camera datasets after cross-testing on one another crowd videos. The robustness and correctness of both entities are contentious, which could be observed in the following cases:

1. The PeopleNet generates promising results at pixel and image level over moving crowd scenes, which are are complementary to each other. Whether the constructed dataset fulfills the current CC research expectations? It is a question.
2. Table III, V and VI validates the correctness of data samples by exploring the model's capability in different crowd scenarios. Whether this novel framework is acceptable for public and standard crowd datasets? It is also a question.

We have answered the above questions in detailed along-with significant case studies 1 and 2.

1. Case Study 1

Testing the crowd counting ability of a novel model on a newly constructed dataset has always been challenging. Comparing the efficiency of some extant object counting models with the proposed methodology is an integral part of any comparative analysis research has been illustrated in Table V, which overviews the comparative performance analysis of existing CC models and the proposed model. As per our best knowledge, due scarcity of tracking camera surveillance crowd datasets, we are forced to compare the Constructed dataset for validation. Some of the universally agreed performance evaluation metrics have been computed for popular deep learning models such as ResNet50 [41], CSRNet [43], DENet [51] and, People-Flow [52] and presented in tabular form. These models are originally developed to perform the crowd counting in free or surveillance view crowd datasets captured by a static camera. Therefore, the performance of these CC models will deteriorate with the proposed database. A PSNR and SSIM values of 26 and 0.43 respectively ensure the high quality of density maps results in accurate people estimation. However, the correlation (C_r) and Pearson coefficient (P_r) variation were observed significantly across all samples. These variations have been received with different samples due to the architectural complexity of various models.

A dilated-CNN structure with 2,160,000 trainable parameters (290.024 MB) in CSRNet [43], is specially designed for density estimation on highly congested crowd datasets such as Shanghai-Tech [40], *WorldExpo* [41], *UCF_CC_50* [56] and UCSD [57] datasets. Its lightweight functionality along with the front (basic CNN) and back-end network is a significant advantage over existing models to generate high-quality density estimation with less hardware computational training effort. A decremented variation of 5-18% and 12-21% for MAE and RMSE has registered over four different random sub-samples, and a decremented variation of 40-75% has registered for

TABLE V. Summary of Comparing Performance Metrics of Different Frameworks With Overall and Segmented Portions (Viz Subset1, Subset2, Etc.) of Self-Constructed Datasets. The Proposed Framework PeopleNet Has Marked With . CSRNet [43] & DENet [51] Frameworks Are Tested Over Two Data Subsets Due to Having Heavily Loaded Front End and Back-End Integrated Networks. The Down Arrow (↓) Means the Lower the Metrics Higher the Accuracy, and the up Arrow (↑) Means the Higher the Metrics Higher the Accuracy. The Term 'NA' Denotes Not Available

Category		**SampleCount**	**PerformanceMetrics**								
			ImageLevel						**PixelLevel**		
Models	**Test Subsets**	**(20:80)/100**	**MAE(↓)**	**RMSE(↓)**	**MAPE(%)(↓)**	**PearsonCoefficient (Pr)(↑)**	**CorrelationCoefficient (r)(↑)**	**R^2_A(↑)**	**nrmse(t)**	**PSNR(dB)(↑)**	**SSIM(↑)**
CSRNet[43]	**Subset1**	110:441/551	17.89	22.43	NA	NA	NA	NA	0.133	13.21	NA
	Subset2	110:441/551	15.51	20.89	NA	NA	NA	NA	0.126	14.76	NA
	Overall	**220:881/1101**	**10.03**	**16.09**	**NA**	**0.922**	**0.931**	**0.941**	**0.117**	**25.99**	**0.43**
ResNet50[41]	**Subset1**	55:220/275	3.97	6.70	18.47	0.953	0.976	0.969	0.142	11.45	NA
	Subset2	55:220/275	3.64	4.89	21.70	0.886	0.942	0.910	0.199	9.67	NA
	Subset3	55:220/275	3.38	5.08	20.21	0.865	0.915	0.892	0.191	5.22	NA
	Subset4	55:220/275	3.38	4.28	21.64	0.856	0.925	0.877	0.181	9.22	NA
	Overall	**220:881/1101**	**4.14**	**5.49**	**17.53**	**0.888**	**0.934**	**0.903**	**0.176**	**19.87**	**0.30**
DENet [51]	**Subset1**	110:441/551	6.60	11.27	21.02	0.887	0.901	0.899	0.155	NA	NA
	Subset2	110:440/551	5.97	11.74	19.83	0.891	0.926	0.917	0.143	NA	NA
	Overall	**220:881/1101**	**4.58**	**6.12**	**18.78**	**0.788**	**0.947**	**0.805**	**0.152**	**NA**	**NA**
FlowNet[52]	**Overall**	**220:881/1101**	**3.23**	**5.33**	**17.25**	**0.81**	**0.905**	**0.822**	**0.152**	**NA**	**0.26**
***PeopleNet* ***	**Subset1**	55:220/275	3.76	5.23	25.10	0.974	0.987	0.988	0.124	39.61	NA
	Subset2	55:220/275	2.95	4.20	15.90	0.902	0.950	0.92	0.143	34.23	NA
	Subset3	55:220/275	3.25	3.45	17.90	0.922	0.932	0.941	0.157	31.21	NA
	Subset4	55:220/275	2.76	3.67	19.03	0.853	0.924	0.862	0.164	47.01	NA
	Overall	**220:881/1101**	**3.306**	**4.38**	**19.75**	**0.977**	**0.954**	**0.923**	**0.160**	**24.12**	**0.52**

TABLE VI. Results Validation on Various Datasets, Highlighted Black Text Represents the Results Obtained on Random Sub-Samples, Highlighted Blue Text Represents the Results Obtained by the PeopleNet Framework

Model	Dataset/ TestSubset		SampleSize (Test:Train)/Total	Performance	
				MAE	RMSE
*PeopleNet**	Mall[49]	Subset-1	100:400/500	1.648	2.116
		Subset-2	100:400/500	1.799	2.193
		Subset-3	100:400/500	1.407	1.708
		Subset-4	100:400/500	1.589	2.094
		Overall	**400:1600/2000**	**1.247**	**1.33**
SAAN[53]			**400:1600/2000**	**1.28**	**1.68**
*PeopleNet**	Beijing-BRT[54]	Subset-1	128:512/640	11.65	16.179
		Subset-2	128:512/640	8.933	14.325
		Overall	**256:1024/1280**	**3.172**	**4.1634**
DRResNet[54]			**256:1024/1280**	**1.39**	**2.00**
*PeopleNet**	Shanghai-Tech-B[40]	Subset-1	71:287/358	20.916	30.015
		Subset-2	71:287/358	25.583	42.108
		Overall	**143:573/716**	**9.807**	**14.265**
SPANet[55]			**143:573/716**	**6.50**	**9.90**

the pixel level accuracy of generated DMs. A PSNR and SSIM values of 26 and 0.43 ensure the high quality of density maps responsible for accurate people estimation.

The ResNet-50 [41] model was tested over four random sub-data samples and secured a decrease of 55-80%, and a decrease of 30-40% in MAE and RMSE respectively. However, an increase of 40-50% has been observed for image level (PSNR) accuracy. This ensures the residual learning and strong feature extraction ability of the standard VGG19 which has been used as a backbone across 16,263,489 trainable parameters of ResNet50. The significant disadvantage of this network is to produce the low quality of density maps with poor regularization techniques, and as results secure lower PSNR and SSIM values. The percentage variation for image and pixel level accuracy is caused by training over lower training samples having mixing data (sparse and dense crowd). However, the lightweight training feature extraction ability is one of the significant advantages of this model, requiring fewer neuron weights (after training 187 MB) to perform pedestrian detection for dense crowd images.

The People-Flow model [52] was used to perform pedestrian estimation via using standard VGG16. It has a front end and its layers as the back end to ensure high-quality density maps. This integrated architecture secures 3.2, and 5.3 as MAE and RMSE; which is good enough but 0.152 as nrmse ensures the model's robustness towards object head courting even in high crowd density samples while securing 0.26 as the SSIM value. The DENet [51] is an integrated detection (DENet) and estimation network (ENet) that performs CC tasks by using an encoder-decoder network as a dual-end network that trains separately over Mask R-CNN (9,64,983 trainable parameters). In comparison to 1101 samples for DENet, an increase of 30-44% in MAE and 85-95% in RMSE is obtained over two random sub-datasets; this shows vibrant performance for data variation. Due to the encoder-decoder architecture of DENet, it disables pixel annotations, and computing the PSNR coefficient is nearly impossible. This universal network achieved MAE and RMSE values of 4.23 and 5.67, respectively, but failed to generate high-quality DMs due to low crowd density.

As we can see in Table V, MAE, RMSE, and MAPE lower values support higher regression accuracy between actual and observed values. The MAPE has calculated by dividing the difference by the actual value, in which if the actual value is close to 0 then the error will be very high, so MAPE is to be used only when the actual value is far from 0. That is the main reason for securing the MAPE of the PeopleNet model is higher than FlowNet for complete data samples. The Pearson and correlation coefficients determine the relationship strength, the higher the value the stronger the relationship. The CSRNet, ResNet50, DeNet, and FlowNet architecture showed a moderate degree of correlation as securing the average Pearson and correlation coefficients over their subset and overall testing samples for crowd counting purposes. Always having a higher correlation never means a strong relationship as it is a bi-variate relationship that somehow depends on the network architecture also, which has observed in the case of subsets testing of CSRNet model rather than PeopleNet. The CSRNet model adds a high number of useful variables as compared to existing models which results in high adjusted R^2. However, the proposed PeopleNet model secures R^2 a total of 0.977 by clearing the misconception of lower regression accuracy for low to adjust R_A^2 instead of having fewer neurons as CSRNet. The nrmse indicator is not always reliable for finding the best networks for small training samples, which is indeed shown in the case of ResNet50. The proposed model traced higher 'nrmse' among the models utilizing front and back-end training architecture. On another hand, the pixel-level performance of the PeopleNet model is incomparable as compared to existing state of art crowd counting models. The better overall object counting accuracy of the proposed model ensures its scalability and robustness even in videos shot in extreme conditions.

2. Case Study 2

Table VI shows the acquired findings over the benchmark and public datasets to compare overall performance without bias. Various statistical reports and results were presented to differentiate the accuracy variance for different data subsets. The RMSE for each random subset having 500 samples in the Mall dataset fluctuates between 28 and 65% when compared to the entire data samples; nevertheless, SAAN [53] achieved a nearly 26% increase.

The sparse, dense sub-sampling disparity has shrunk marginally in MAE, but it is essentially non-existent in SAAN [53] and PeopleNet. Because the Beijing-BRT [54] includes a total of 1280 samples, we computed PeopleNet results for two random and equal subsets due to the smaller data samples. We can see the nuanced variance in MAE but not in RMSE in each case; however, the proposed model has demonstrated superiority over the DRResNet [54] model by obtaining near 50% reductions in MAE and RMSE. There are 716 samples in the ShanghaiTech-B [40]. As a result, we only assessed PeopleNet's performance on two random samples, subset1 and subset2. For the Beijing-BRT and ShanghaiTech-B samples, there was a 100-200% change in MAE and RMSE, with a smaller percentage loss for DRResNet [54] and SPANet [55] accuracy. The obtained results present some image-level performance on frequently used databases; however, the pixel-level accuracy comparison is not helpful for static devices captured in existing datasets. The obtained results by the PeopleNet model on the proposed dataset are closer to the obtained on existing datasets; which is solid evidence of the correctness, scalability, and robustness of the proposed model over constructed data samples.

V. Conclusion

Employing a novel PeopleNet framework, this study handled a difficult CV problem of autonomous crowd counting using a tracking dynamic imager. Using the feature of expansion residual mapping over the camera-induced motion for a self-generated head-mounted video dataset, this mechanism performs BEYOND and IN operations for visible spectrum. The technical aspect of this model is to provide fair people counting over moving cameras and moving people without intentionally or unintentionally pointing out individuals. The behavioral aspect of this study includes human counting in dangerous enemy territory or isolated places where electricity and infrastructure are no longer available. The PeopleNet's experimental findings revealed that pedestrian recognition is done efficiently in the day or night environments to address occlusion. Detection of social distance practice violation via crowd density estimation could be another significant social aspect of this work.

Extend the PeopleNet model's functionality for Covid19 like virus protocols via crowd monitoring in public places will be the possible future scope of this work.

Acknowledgment

We are grateful to Graphic Era University to provide the computational resources execution of this research work. We also give special thanks to the research team and websites for providing the public Mall, SmartCity, and ShanghaiTech-B dataset repositories. We are also grateful to the anonymous reviewers for providing us with valuable comments to make this research work better.

References

[1] A. Ferligoj, V. Batagelj, "Direct multicriteria clustering algorithms," *Journal of classification*, vol. 9, no. 1, pp. 43– 61, 1992.

[2] H. Faris, I. Aljarah, S. Mirjalili, "Training feedforward neural networks using multi-verse optimizer for binary classification problems," *Applied Intelligence*, vol. 45, pp. 322–332, 2016.

[3] A. Korotayev, J. Zinkina, "Egypt's 2011 revolution: A demographic structural analysis," in *Handbook of revolutions in the 21st century: The new waves of revolutions, and the causes and effects of disruptive political change*, Springer, 2022, pp. 651–683.

[4] C. A. Martin, C. Marshall, P. Patel, C. Goss, D. R. Jenkins, C. Ellwood, L. Barton, A. Price, N. J. Brunskill, K. Khunti, *et al.*, "Association of demographic and occupational factors with sars-cov-2 vaccine uptake in a multi-ethnic uk healthcare workforce: a rapid real- world analysis," *MedRXiv*, pp. 2021–02, 2021.

[5] E. A. Felemban, F. U. Rehman, S. A. A. Biabani, A. Ahmad, A. Naseer, A. R. M. A. Majid, O. K. Hussain, A. M. Qamar, R. Falemban, F. Zanjir, "Digital revolution for hajj crowd management: a technology survey," *IEEE Access*, vol. 8, pp. 208583–208609, 2020.

[6] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, *et al.*, "A system for video surveillance and monitoring," *VSAM final report*, vol. 2000, no. 1-68, p. 1, 2000.

[7] M. Adimoolam, S. Mohan, G. Srivastava, *et al.*, "A novel technique to detect and track multiple objects in dynamic video surveillance systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, pp. 112–120, 2022.

[8] A. Sobral, E.-h. Zahzah, "Matrix and tensor completion algorithms for background model initialization: A comparative evaluation," *Pattern Recognition Letters*, vol. 96, pp. 22–33, 2017.

[9] B. Xu, G. Qiu, "Crowd density estimation based on rich features and random projection forest," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–8, IEEE.

[10] C. Arteta, V. Lempitsky, J. A. Noble, A. Zisserman, "Interactive object counting," in *Computer Vision– ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, 2014, pp. 504–518, Springer.

[11] C. Zhang, Z. Liu, C. Bi, S. Chang, "Dependent motion segmentation in moving camera videos: A survey," *IEEE Access*, vol. 6, pp. 55963–55975, 2018.

[12] D. Wang, "Unsupervised video segmentation based on watersheds and temporal tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 539–546, 1998.

[13] D. Comaniciu, P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[14] A. Ghazvini, S. N. H. S. Abdullah, M. Ayob, "A recent trend in individual counting approach using deep network," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, pp. 7–14, 2019.

[15] A. Ghosh, B. N. Subudhi, S. Ghosh, "Object detection from videos captured by moving camera by fuzzy edge incorporated markov random field and local histogram matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 8, pp. 1127–1135, 2012.

[16] P.-M. Jodoin, M. Mignotte, C. Rosenberger, "Segmentation framework based on label field fusion," *IEEE Transactions on Image Processing*, vol. 16, no. 10, pp. 2535–2550, 2007.

[17] B. Lee, M. Hedley, "Background estimation for video surveillance," in *Image and Vision Computing*, 2002, pp. 315–320, CSIRO.

[18] C. Stauffer, W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition*, vol. 2, 1999, pp. 246–252, IEEE.

[19] O. Munteanu, T. Bouwmans, E. Zahzah, R. Vasiu, "The detection of moving objects in video by background subtraction using dempster-shafer theory," *Transactions on Electronics and Communications*, vol. 60, no. 1, pp. 1–9, 2015.

[20] C. Marghes, T. Bouwmans, R. Vasiu, "Background modeling and foreground detection via a reconstructive and discriminative subspace learning approach," in *International Conference on Image Processing, Computer Vision, and Pattern Recognition, IPCV*, vol. 2012, 2012.

[21] R. C. Joshi, A. G. Singh, M. Joshi, S. Mathur, "A low-cost and computationally efficient approach for occlusion handling in video surveillance systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, pp. 28–38, 2019.

[22] J. A. Ramirez-Quintana, M. I. Chacon-Murguia, "Self- adaptive som-cnn neural system for dynamic object detection in normal and complex scenarios," *Pattern Recognition*, vol. 48, no. 4, pp. 1137–1149, 2015.

[23] J. A. Ramírez-Quintana, M. I. Chacon-Murguía, "Self- organizing retinotopic maps applied to background modeling for dynamic object segmentation in video sequences," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 2013, pp. 1–8, IEEE.

[24] C. R. Wren, A. Azarbayejani, T. Darrell, A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 780–785, 1997.

[25] H.-x. Zhang, D. Xu, "Fusing color and gradient features for background model," in *2006 8th International Conference on Signal Processing*, vol. 2, 2006, IEEE.

[26] D. Zeng, M. Zhu, A. Kuijper, "Combining background subtraction algorithms with convolutional neural network," *Journal of Electronic Imaging*, vol. 28, no. 1, pp. 013011–013011, 2019.

[27] M. Babaee, D. T. Dinh, G. Rigoll, "A deep convolutional neural network for background subtraction," *arXiv preprint arXiv:1702.01731*, 2017.

[28] L. Xu, Y. Li, Y. Wang, E. Chen, "Temporally adaptive restricted boltzmann machine for background modeling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, 2015.

[29] M. J. Shafiee, P. Siva, P. Fieguth, A. Wong, "Real- time embedded motion detection via neural response mixture modeling," *Journal of Signal Processing Systems*, vol. 90, pp. 931–946, 2018.

[30] M. J. Shafiee, P. Siva, P. Fieguth, A. Wong, "Embedded motion detection via neural response mixture background modeling," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops*

(CVPRW), 2016, pp. 837–844, IEEE.

[31] R. Guo, H. Qi, "Partially-sparse restricted boltzmann machine for background modeling and subtraction," in *2013 12th International Conference on Machine Learning and Applications*, vol. 1, 2013, pp. 209–214, IEEE.

[32] P. Xu, M. Ye, X. Li, Q. Liu, Y. Yang, J. Ding, "Dynamic background learning through deep auto- encoder networks," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 107– 116.

[33] Z. Qu, S. Yu, M. Fu, "Motion background modeling based on context-encoder," in *2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR)*, 2016, pp. 1–5, IEEE.

[34] M. Braham, M. Van Droogenbroeck, "Deep background subtraction with scene-specific convolutional neural networks," in *2016 international conference on systems, signals and image processing (IWSSIP)*, 2016, pp. 1–4, IEEE.

[35] L. P. Cinelli, "Anomaly detection in surveillance videos using deep residual networks," *Universidade Federal do Rio de Janeiro, Rio de Janeiro*, 2017.

[36] C. M. Bautista, C. A. Dy, M. I. Mañalac, R. A. Orbe, M. Cordel, "Convolutional neural network for vehicle detection in low-resolution traffic videos," in *2016 IEEE Region 10 Symposium (TENSYMP)*, 2016, pp. 277–281, IEEE.

[37] X. Zhao, Y. Chen, M. Tang, J. Wang, "Joint background reconstruction and foreground segmentation via a two-stage convolutional neural network," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 2017, pp. 343–348, IEEE.

[38] J. Wang, K. L. Chan, "Background subtraction based on encoder-decoder structured cnn," in *Pattern Recognition: 5th Asian Conference, ACPR 2019, Auckland, New Zealand, November 26–29, 2019, Revised Selected Papers, Part II 5*, 2020, pp. 351–361, Springer.

[39] Y. Wang, Z. Luo, P.-M. Jodoin, "Interactive deep learning method for segmenting moving objects," *Pattern Recognition Letters*, vol. 96, pp. 66–75, 2017.

[40] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, "Single- image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 589– 597.

[41] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[42] L. Boominathan, S. S. Kruthiventi, R. V. Babu, "Crowdnet: A deep convolutional network for dense crowd counting," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 640– 644.

[43] Y. Li, X. Zhang, D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1091–1100.

[44] G. Gao, J. Gao, Q. Liu, Q. Wang, Y. Wang, "Cnn-based density estimation and crowd counting: A survey," *arXiv preprint arXiv:2003.12783*, 2020.

[45] V. Lempitsky, A. Zisserman, "Learning to count objects in images," *Advances in neural information processing systems*, vol. 23, 2010.

[46] S. Kumagai, K. Hotta, T. Kurita, "Mixture of counting cnns," *Machine Vision and Applications*, vol. 29, no. 7, pp. 1119–1126, 2018.

[47] Y. Zhang, S. J. Kiselewich, W. A. Bauson, R. Hammoud, "Robust moving object detection at distance in the visible spectrum and beyond using a moving camera," in *2006 conference on computer vision and pattern recognition workshop (CVPRW'06)*, 2006, pp. 131–131, IEEE.

[48] K. K. Verma, B. M. Singh, "Deep multi-model fusion for human activity recognition using evolutionary algorithms," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, pp. 44– 58, 2021.

[49] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, S. Yan, "Crowded scene analysis: A survey," *IEEE transactions on circuits and systems for video technology*, vol. 25, no. 3, pp. 367–386, 2014.

[50] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[51] L. Liu, J. Jiang, W. Jia, S. Amirgholipour, Y. Wang, M. Zeibots, X. He, "Denet: A universal network for counting crowd with varying densities and scales," *IEEE Transactions on Multimedia*, vol. 23, pp. 1060–1068, 2020.

[52] L. Liu, H. Lu, H. Zou, H. Xiong, Z. Cao, C. Shen, "Weighing counts: Sequential crowd counting by reinforcement learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, 2020, pp. 164–181, Springer.

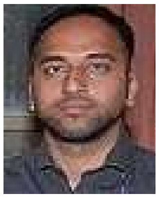
[53] M. Hossain, M. Hosseinzadeh, O. Chanda, Y. Wang, "Crowd counting using scale-aware attention networks," in *2019 IEEE winter conference on applications of computer vision (WACV)*, 2019, pp. 1280–1288, IEEE.

[54] X. Ding, Z. Lin, F. He, Y. Wang, Y. Huang, "A deeply- recursive convolutional network for crowd counting," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 1942–1946, IEEE.

[55] Z.-Q. Cheng, J.-X. Li, Q. Dai, X. Wu, A. G. Hauptmann, "Learning spatial awareness to improve crowd counting," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6152–6161.

[56] H. Idrees, I. Saleemi, C. Seibert, M. Shah, "Multi- source multi-scale counting in extremely dense crowd images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2547–2554.

[57] A. B. Chan, Z.-S. J. Liang, N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *2008 IEEE conference on computer vision and pattern recognition*, 2008, pp. 1–7, IEEE.



Ankit Tomar

Mr. Ankit Tomar has done B Tech Computer Science and Engineering from UPTU, M-Tech from Jamia Hamdard, New Delhi, and pursuing a Ph.D. from Graphic Era Deemed University, Dehradun. He is actively involved in research related to Deep Learning and Machine Learning. He has published many papers in reputed international conferences and journals. Currently, he is working as an assistant professor in the Graphic era deemed by the university Dehradun, India.

Santosh Kumar

Dr. Santosh Kumar earned his Ph.D. from India Institute of Technology, Roorkee, India in 2012, M. Tech. in Computer Science and Engineering from Aligarh Muslim University, Aligarh, India in 2007, and B.E. (IT) from C.C.S. University, Meerut, India in 2003. He is an active reviewer board member in various national/International Journals and Conferences. He has memberships of ACM (Senior Member), IAENG, ACEEE, and ISOC (USA) and contributed more than 80 research papers in National and International Journals/conferences. Currently holding a position of Professor in the Graphic Era Deemed to be University, Dehradun, India. His research interest includes AI &; Machine Learning, Wireless Networks, WSN, IoT, and Software Engineering.

Bhaskar Pant

Dr. Bhaskar Pant is currently working as Dean of Research & Development and Associate Professor in the Department of Computer Science and Engineering. He is Ph.D. in Machine Learning and Bioinformatics from MANIT, Bhopal. Has more than 17 years of experience in Research and Academics. He has till now guided as Supervisor 4 Ph.D. candidates (Awarded).and 7 candidates are in an advanced state of work. He has also guided 32 M. Tech. Students for dissertation. He has also supervised 2 foreign students for internship. Dr. Bhasker Pant has more than 100 research publications in National and international Journals.

Lightweight Real-Time Recurrent Models for Speech Enhancement and Automatic Speech Recognition

Sami Dhahbi[1], Nasir Saleem[2]*, Teddy Surya Gunawan[3], Sami Bourouis[4], Imad Ali[5], Aymen Trigui[6], Abeer D. Algarni[7]

[1] Department of Computer science, College of science and art at Mahayil, King Khalid University, Muhayil Aseer, 62529 (Saudi Arabia)
[2] Department of Electrical Engineering, FET, Gomal University, D.I. Khan-29050, KPK (Pakistan)
[3] Electrical and Computer Engineering Department, Islamic International University Malaysia, Kuala Lumpur (Malaysia)
[4] Department of Information Technology, College of Computers and Information Technology, Taif University, Taif 21944 (Saudi Arabia)
[5] Department of Computer Science, University of Swat, Swat (Pakistan)
[6] Department of Computer Science, College of Computer Science, King Khalid University, Abha (Saudi Arabia)
[7] Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671 (Saudi Arabia)

* Corresponding author. nasirsaleem@gu.edu.pk

Received 15 November 2023 | Accepted 30 January 2024 | Early Access 30 April 2024

Abstract

Traditional recurrent neural networks (RNNs) encounter difficulty in capturing long-term temporal dependencies. However, lightweight recurrent models for speech enhancement are important to improve noisy speech, while being computationally efficient and able to capture long-term temporal dependencies efficiently. This study proposes a lightweight hourglass-shaped model for speech enhancement (SE) and automatic speech recognition (ASR). Simple recurrent units (SRU) with skip connections are implemented where attention gates are added to the skip connections, highlighting the important features and spectral regions. The model operates without relying on future information that is well-suited for real-time processing. Combined acoustic features and two training objectives are estimated. Experimental evaluations using the short time speech intelligibility (STOI), perceptual evaluation of speech quality (PESQ), and word error rates (WERs) indicate better intelligibility, perceptual quality, and word recognition rates. The composite measures further confirm the performance of residual noise and speech distortion. With the TIMIT database, the proposed model improves the STOI and PESQ by 16.21% and 0.69 (31.1%) whereas with the LibriSpeech database, the model improves STOI by 16.41% and PESQ by 0.71 (32.9%) over the noisy speech. Further, our model outperforms other deep neural networks (DNNs) in seen and unseen conditions. The ASR performance is measured using the Kaldi toolkit and achieves 15.13% WERs in noisy backgrounds.

Keywords

Real-Time Speech, Simple Recurrent Unit (SRU), Speech Enhancement, Speech Processing, Speech Quality.

DOI: 10.9781/ijimai.2024.04.003

I. Introduction

Building lightweight recurrent models for speech enhancement and Automatic Speech Recognition (ASR) involves designing models that can process audio data efficiently while improving speech quality and intelligibility, mostly in noisy or degraded environments. The speech enhancement (SE) is particularly significant as it reduces listener fatigue, especially when individuals are subjected to prolonged exposure to high noise. The SE positively impacts the efficiency of communication and multimedia systems. Furthermore, it improves the intelligibility of speech, thereby enhancing ASRs and interactions between humans and machines. Various proposals are available in the literature, encompassing methods such as spectral subtraction [1]-[2], Wiener filtering [3], and minimum mean square error (MMSE) [4]-[5].

To address the SE challenges, supervised learning models are considered. These models undergo training using large speech datasets [6]-[7]. Among successful models for SE are the regression-based deep neural networks (DNNs) [8]-[11]. Given that the relationship between input and target features is nonlinear, a multi-layer DNN incorporating nonlinear activation is a suitable option. Essential considerations for a DNN-based SE include the type of network, the learning objective, and

Please cite this article as: S. Dhahbi, N. Saleem, T. S. Gunawan, S. Bourouis, I. Ali, A. Trigui, A. D. Algarni, “Lightweight Real-Time Recurrent Models for Speech Enhancement and Automatic Speech Recognition”, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 6, pp. 74-85, 2024, http://dx.doi.org/10.9781/ijimai.2024.04.003

the loss function [12]-[14]. For SE, the learning models are categorized into spectral mapping and masking. Mapping models entail training networks through direct mapping rules. They learn to estimate clean spectral features from noisy spectral features. However, these methods often result in excessively smooth spectra [9]. Conversely, masking-based learning algorithms have shown greater success in SE. They involve multiplying the estimated parameters of objectives (ideal ratio mask (IRM) or ideal binary mask (IBM)) with noisy magnitudes. Many deep-learning approaches have recently emerged time-frequency (T-F) masks as training objectives, yielding favorable results [15]-[21]. Fully connected feedforward DNNs (FDNNs) predict labels for individual time frames using small context windows. Yet, they lack control over the long-term context windows crucial for accurately tracking the target speaker. DNN-based SE algorithms employ multi-layer DNNs for learning nonlinear regression functions or estimate a spectral mask using noisy magnitudes. These models forego the requisite for statistical distributions assumption, yielding superior noise reduction when handling non-stationary noises. The SE system using recurrent neural networks (RNNs) with T-F masking is depicted in Fig. 1.

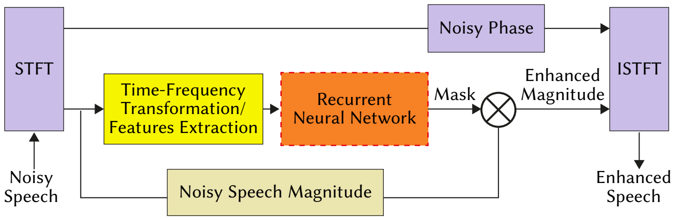


Fig. 1. SE with RNNs estimating T-F masking.

Recurrent neural network (RNN) has attainment significance in several challenging applications within Natural Language Processing, including neural machine translation [22], conversational/dialogue modeling [23], and ASR modeling [24]. Given that speech waveform is a sequential data type, it requires a temporal context for effective processing, and RNNs excel in capturing long-range temporal sequences. Previous research [10] [25] has recommended framing SE as a sequence-to-sequence procedure to manage long-term contextual window. Various models, including RNN, CNN [57], and other deep learning architectures, have been proposed and assessed on diverse noises and speakers. In a work [25], LSTM is introduced for speaker's generalization. The results indicate that the LSTM model demonstrates superior generalization to untrained speakers, significantly outperforming a DNN-based model in terms of speech intelligibility. Multiple studies have emphasized that employing the sequence-to-sequence approach enables LSTM to effectively control long-term context windows, leading to successful outcomes in speech enhancement [26]. To model the long input sequential data, RNNs face problems in capturing long-term temporal dependencies. Further, training an RNN with back propagation through time is exposed to vanish and explode the gradients. These challenges are addressed by proposing RNN variants using novel transition functional units and optimization techniques, such as LSTM [11] [27] and gated recurrent unit (GRU) [28]-[29]. Several approaches focused on connection architectures, including stacked RNNs [30] and skip RNNs [31]. In this paper, we have proposed efficient simple recurrent unit (SRU) models that are able to detain the long-term temporal dependencies and prevent the gradient from decaying. The contributions are highlighted below.

- A proposed SRU model takes on an hourglass shape, effectively capturing long-term temporal and sequential data. This results in reduced feature resolutions without sacrificing data in the layers.
- Skip connections are introduced between nonadjacent layers to mitigate decaying gradient. Additionally, attention gates within the skip connections are used to reduce irrelevant features and highlight crucial features across different spectral regions.
- Robust training of the proposed SRU-based model is achieved by extracting combined feature sets from the noisy speech.
- We estimate two distinct training objectives, Ideal Ratio Mask and Ideal Binary Mask, to attenuate noise in the noisy mixture. This approach aims to enhance speech quality, intelligibility, and reduce word error rates.

The rest of this study is structured as follows: Section II outlines the formulation of the SE problem. Section III introduces the proposed SE. Details of the experiments conducted are outlined in Section IV. Section V provides the results and discussions. Ultimately, Section VI presents the drawn conclusions.

II. Problem Formulation

Take into account that a clear speech signal $x(t)$ undergoes degradation due to presence of background noise $n(t)$. This leads to the generation of a noisy speech signal $s(t)$, which can be represented as:

$$s(t) = x(t) + n(t) \tag{1}$$

The noisy speech signal, denoted as $s(t)$, undergoes a transformation to the frequency domain through application of the short-time Fourier Transform (STFT). This results in the acquisition of the frequency domain depiction:

$$|S(f,t)| = |X(f,t)| + |N(f,t)| \tag{2}$$

Where t and f denote the frame and frequency indexes, respectively. A combined feature set is extracted to robustly train the proposed SRU model. During inference, the trained parameters estimate Ideal Ratio Mask and the Ideal Binary Mask as training objectives. The estimated magnitude masks are then multiplied by noisy magnitudes to suppress the background noises:

$$|\hat{X}(f,t)| = |M_x(f,t)| \otimes |S(f,t)| \tag{3}$$

Where |M(t, f)| is the estimated T-F mask. The estimated magnitude and noisy phase reconstruct noise-free enhanced speech waveforms. The block diagram of the proposed SE is depicted in Fig. 2.

III. Proposed Speech Enhancement

SRUs can detain the information in speech waveforms which is a kind of long-term temporal sequence. The proposed network architecture effectively addresses the limitations of traditional RNNs using the following approaches. Our aim is to reduce the complexity (which is directly linked with neurons quantity and number of steps) without degrading the speech enhancement performance. Since equal number of neurons is each layer will introduce computational load, we have arranged neurons in increasing-decreasing order which forms a U-Shape layer. With this arrangement, the overall complexity of the model is reduced (with reduced neurons). Further, the same mechanism is adopted for time steps. By increasing the number of time steps, the computational complexity can indeed increase. This is because the model needs to process information across multiple time steps, leading to a higher demand for computational resources. More time steps may require more complex models to capture long-term dependencies in the data. Firstly, the network architecture has a unique shape with bottom and upper pyramids. For the upper pyramid, there is a decrease in time steps while the number of neurons increases. Conversely, the lower pyramid exhibits an increase in time steps paired with a decrease in number of neurons across layers. This architectural design enables the model to manage high-resolution

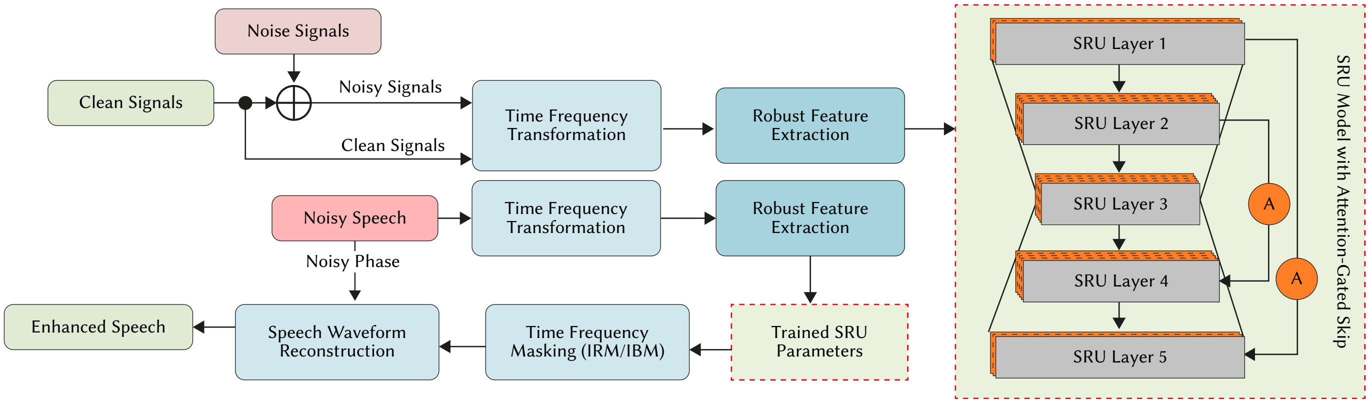

Noise Signals
Clean Signals
Noisy Signals
Clean Signals
Time Frequency Transformation
Robust Feature Extraction
Noisy Speech
Noisy Phase
Time Frequency Transformation
Robust Feature Extraction
Enhanced Speech
Speech Waveform Reconstruction
Time Frequency Masking (IRM/IBM)
Trained SRU Parameters
SRU Layer 1
SRU Layer 2
SRU Layer 3
SRU Layer 4
SRU Layer 5
A
A
SRU Model with Attention-Gated Skip

Fig. 2. The block diagram of the proposed Speech Enhancement System.

features without encountering memory overflow issues. Additionally, skip connections have been incorporated between layers of similar shapes, spanning from upper to lower pyramid. This inclusion enhances the mitigation of gradient decay throughout the layers. To further refine the skips, an attention gate is introduced, emphasizing crucial spectral regions. Given that the speech spectrum showcases dominant formants in low-frequency areas and a sparse distribution in high-frequency zones, it becomes imperative to employ attention weights for distinguishing these varied spectral regions through the attention process. The model consists of five SRU layers featuring two attention-gated skip connections, as illustrated in Fig. 3. Details regarding time steps and units can be found in Table I. The network adeptly learns the nonlinear relationship, effecting the transformation of noisy speech, denoted as *s*(t), into a clear and intelligible speech signal, denoted as *x*(t).

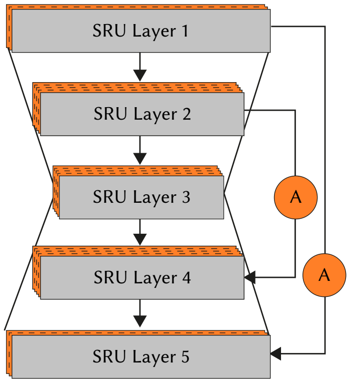


Fig. 3. The proposed SRU architecture.

TABLE I. Proposed SRU Details

Layer	Neurons in Layer	Time-Steps in Layers
1	256	512
2	512	256
3	1024	128
4	512	256
5	256	512

The SRU is a type of recurrent neural network (RNN) that incorporates parallelism analogous to convolutional and feedforward networks. It achieves an exact stability between sequential dependency and independency. While the state computation in SRU is time-dependent, each dimension within the state operates independently. Additionally, SRU improves the training of deep recurrent models through the integration of highway connections and a specialized parameter initialization technique tailored for effective gradient propagation in deep architectures. The computational process of SRU includes the following steps.

$$f_t = \sigma(W_f x_t + b_f) \tag{4}$$

$$r_t = \sigma(W_r x_r + b_r) \tag{5}$$

$$C_t = f_t \otimes C_{t-1} + (1 - f_t) \otimes W x_t \tag{6}$$

$$h_t = r_t \otimes g C_t + (1 - r_t) \otimes x_t \tag{7}$$

Where **W** and **b** are learnable weight matrices and bias terms. In the SRU architecture, the computation of forget and reset gates is independent, eliminating interdependence and simplifying the gating mechanism for faster training. Moreover, the candidate hidden state is determined through element-wise multiplication of the reset gate with the previous hidden state, enhancing SRU's ability to capture long-term dependencies more effectively compared to the traditional RNN. This streamlined SRU design enhances network capacity by sharing hidden states among similar and lower layers. The SRU architecture incorporates a strategy of decreasing time-steps and increasing neurons from the first layer to the mid-layer, and conversely, increasing time-steps and decreasing neurons from the mid-layer to the last layer. This approach facilitates a deeper representation. The shared hidden states among layers in SRU mean that the hidden states in layer '*l*' at time '*t*' are derived by combining hidden states from the (*l*-1) lower layer at time (*t*-1). Before the skips, the hidden states of upper and lower layers are combined, resulting in a final output similar to the input vectors:

$$h_t^l = SRU(h_t^{l-1}, h_{t-1}^l) \tag{8}$$

The output vector is generated through the combination of the hidden states across all layers, as follows:

$$Z = SRU(h_l^5, h_T^5) \tag{9}$$

Here, Z represents the output vector from the final layer in the SRU. To prevent gradient decay across the layers, two skips are introduced. These skips promote enhanced generalization by integrating low-level and high-level features. Given that speech spectra include various frequency components with formants dominating in the low-frequency regions and displaying sparse distributions in higher-frequency areas, it becomes crucial to employ an attention process that assigns attention weights to discern different frequency regions. This attention process focuses on crucial regions and features, thereby enhancing the output quality. Initially, the alignment vector is calculated for the output yout of the layer as follows:

$$V_{align} = tanh(y_{out} \otimes W) \tag{10}$$

Where **W** indicates trainable weights. The score λ for corresponding alignment vector is given as:

$$\lambda_{score} = \alpha(V_{align} \otimes \beta) \tag{11}$$

The dynamic range for λ is 0 to 1. To avoid weak scores, a controlling parameter β is incorporated. The parameter λ assigns different scores to different features in the feature space. The output of the attention process is given as:

$$\hat{y}out = \gamma_{score} \odot yout \tag{12}$$

Where ⊙ denotes Hadamard product used to weigh all feature streams by using the obtained Scores. The feature-level computation process of attention weights is depicted in Fig. 4. The features are derived from input frames of the speech. The frame shift and duration remain at a consistent 10 milliseconds and 20 milliseconds. The features consist of 31-dimensional Mel-Frequency Cepstral Coefficients, 13-dimensional Relative Spectral Transformed Perceptual Linear Prediction Coefficient, 64-dimensional Gammatone Filter-bank, and 15-dimensional Amplitude Modulation Spectrogram, outlined as follows:

$$f_s = f_s^{MFCC} + f_s^{RASTA-PLP} + f_s^{AMS} + f_s^{GFE} \tag{13}$$

$$f_x = f_x^{MFCC} + f_x^{RASTA-PLP} + f_x^{AMS} + f_x^{GFE} \tag{14}$$

Where f_x and f_s show feature sets of clean x(t) and noisy speech s(t), respectively. GFE features are derived from a Cochleagram [32]. The delta features (Δfx and Δfs) are affixed to the features.

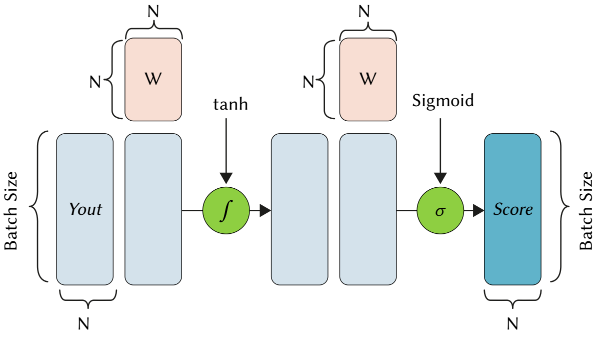


Fig. 4. Attention procedure over Features.

IV. Experiments

A. Dataset

In order to evaluate the proposed SE, a set of experiments is conducted exploiting speech utterances from the TIMIT [33] and LibriSpeech [34] databases. The TIMIT database comprises phonetically balanced speech waveforms sampled at a rate of 16 kHz, while LibriSpeech encompasses 1000 hours of speech waveforms sampled at the rate of 16 kHz. The experiments exclusively utilize clean speech utterances obtained from both databases. For evaluating the effectiveness of the proposed SE under varying noisy conditions, various background noises are selected from Noisex-92 and Aurora-4 databases [35]-[36]. To mix noises with clean utterances, four signal-to-noise ratios (SNRs) are utilized, ranging from -8dB to 4dB in increments of 4dB. To train the SRU network proposed in this study, sentences are selected from the TIMIT database. These utterances are used to generate an ideal ratio mask (IRM) and an ideal binary mask (IBM) for each SNR. To enhance the generalization of the model across different speakers, the training sentences belong to male/female speakers and degraded by all noises. This amounted to 10500 training sentences from TIMIT database. An additional 1500 utterances were randomly chosen for testing. All noise sources, except for two (factory2 and cafeteria), are used in both the training and testing. These two sources are reserved as unseen noises. Furthermore, 2000 sentences are extracted from the LibriSpeech dataset and used to estimate IRM and IBM across all SNRs (8dB to 4dB). This creates a total of 10500 training utterances from the LibriSpeech. In this case, five noise sources are introduced in experiments for training the models with LibriSpeech. These noise sources are airport, babble, street, cafeteria, and car noise, respectively.

B. Network Setting

This study uses five-layered SRU network to enhance speech degraded by noise. The input layer is provided with a 1408-d context window containing 11 frames. Each SRU layer comprises M neurons and N time steps, while the output layer encompasses 257 neurons. During the training process, backpropagation through time (BPTT) [37] is utilized. For optimization, the adaptive gradient descent [38] method with a momentum parameter m is employed, where a scaling factor of 0.0010 is set for AGD. The learning rate follows a linear reduction from 0.06 to 0.002 over the course of processing. Samples of 512 batch size are chosen for training. A total of 80 epochs are completed, during which m remains constant at 0.4 for initial epochs, and subsequently, it is raised to 0.8 for other epochs. Dropout regularization [39] with a 0.02 rate is used to mitigate overfitting. During mask estimation, the Mean Squared Error (MSE) is used as a loss function. Notably, the SRU model operates without exploiting future information, ensuring causality. To estimate current speech frame, a feature context window of 11 frames (comprising 10 previous and 1 current frame) is employed. This approach involves concatenating 11 frames of features into extended vectors, serving as the network's input for each time step, as depicted in Fig. 5. For further details regarding the deep model's hyperparameters, refer to Table II.

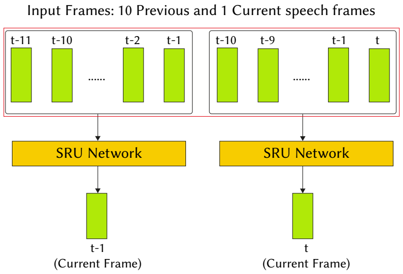


Fig. 5. Causal SRU with feature window of 11 frames.

TABLE II. Proposed SRU Details

Hyper Parameters	Baseline SRU	Baseline DNN	Proposed SRU
Hidden Layers	5	5	5
Layer 1 Neurons	1024	1024	256
Layer 2 Neurons	1024	1024	512
Layer 3 Neurons	1024	1024	1024
Layer 4 Neurons	1024	1024	512
Layer 5 Neurons	1024	1024	256
Learning Rate	0.0001	0.0001	0.0001
No. of Epochs	80	80	80
Momentum Rate	0.8	0.8	0.8
Dropout Rate	0.2	0.2	0.2
Loss Function	MSE	MSE	MSE
Activation	ReLU	ReLU	ReLU

C. Evaluation Metrics

The assessment of our SE involves the use of four objective metrics during the experiments. These metrics encompass the short-time objective intelligibility (STOI), the perceptual evaluation of speech quality (PESQ), and composite measures (CM). These metrics serve purpose of evaluating intelligibility, quality, distortion, and residual noise. The perceptual speech quality, as determined by PESQ [40] following ITU-T P.862 guidelines, is scored within -0.5 to 4.5. STOI [41] quantifies speech intelligibility from 0 to 1 with percentage. Further, the composite measures [42] consist of the C_{SIG} (indicating speech distortions) and the C_{BAK} (reflecting residual noise) [56].

D. Model Representation

To evaluate the proposed SE models, various configurations are considered, each with specific interpretations. The SRU-NoSkip-IRM model estimates IRM using the proposed SRU architecture without skip connections, while the SRU-NoSkip-IBM focuses on estimating IBM without skip connections. In contrast, the SRU-WithSkip-IRM and SRU-WithSkips-IBM models aim to estimate IRM and IBM, respectively, utilizing the proposed SRU with skip connections. Additionally, the SRU-AttSkip-IRM and SRU-AttSkip-IBM models incorporate attention skip connections in their quest to estimate IRM and IBM. The baseline SRU [25], denoted as SRU-IRM and SRU-IBM, employs IRM and IBM as training objectives. All models are trained using the TIMIT and LibriSpeech datasets.

V. Results and Discussions

A. Speech Enhancement in Seen Noises

Table III and Table IV illustrate a comparison of our speech enhancement (SE) algorithms across three distinct noise types, evaluated by STOI. The training objectives involve estimating the Ideal Ratio Mask (IRM) and the Ideal Binary Mask (IBM). The SRU model, incorporating mask estimation, combined feature sets, and attention skips, demonstrated superior performance in comparison to networks lacking skips or utilizing skips without attention. Enhanced intelligibility and quality were observed in the proposed models when applied to noisy speech. For instance, in Table III and Table IV, both SRU-AttSkip-IRM and SRU-AttSkip-IBM exhibited improvements in STOI by 7.7% and 6.9%, respectively, over noisy speech (UNP) at -8dB babble noise. Similarly, at -4dB car noise, these models improved STOI by 23.9% and 23.5%. At 0dB factory noise, the SRU-AttSkip-IRM and SRU-AttSkip-IBM showed STOI improvements of 20.2% and 19.7% over noisy speech. In comparison to the SRU-WithSkip-IRM and SRU-WithSkip-IBM, the proposed models with attention skips achieved a 2.1% and 2.5% improvement in STOI at -8dB babble noise. Additionally, these attention skip models outperformed SRU-NoSkip-IRM and SRU-NoSkip-IRM by 9.1% and 8.5% at -8dB babble noise. Overall, SRU-AttSkip-IRM exhibited notable advantages over SRU-AttSkip-IBM, displaying improved average STOI across noise types and Signal-to-Noise Ratios (SNRs) by 1.23%.

TABLE III. STOI in Seen Noise for IRM Training-Objective

Noise	Model	-8dB	-4dB	0dB	4dB
Babble Noise	Noisy Mixture	48.2	58.1	67.1	76.2
	SRU-NoSkips	52.7	66.7	77.0	84.8
	SRU-WithSkips	53.8	68.8	79.1	87.0
	SRU-AttenSkips	55.9	70.1	80.3	88.6
Car Noise	Noisy Mixture	51.8	58.9	68.6	77.1
	SRU-NoSkips	72.4	79.2	84.9	89.4
	SRU-WithSkips	74.5	86.9	86.9	91.6
	SRU-AttenSkips	75.7	88.3	88.3	93.2
Factory Noise	Noisy Mixture	55.2	61.7	69.8	78.0
	SRU-NoSkips	66.3	76.7	84.5	90.5
	SRU-WithSkips	68.3	78.8	86.5	92.6
	SRU-AttenSkips	69.5	79.9	87.9	93.7

TABLE IV. STOI in Seen Noise for IBM Training-Objective

Noise	Model	-8dB	-4dB	0dB	4dB
Babble Noise	Noisy Mixture	48.2	58.1	67.1	76.2
	SRU-NoSkips	51.6	66.7	77.0	84.8
	SRU-WithSkips	52.6	64.7	77.5	85.7
	SRU-AttenSkips	55.1	66.2	79.8	88.1
Car Noise	Noisy Mixture	51.8	58.9	68.7	77.1
	SRU-NoSkips	72.4	79.2	85.0	89.4
	SRU-WithSkips	71.5	84.3	84.3	88.8
	SRU-AttenSkips	75.3	87.7	87.7	92.2
Factory Noise	Noisy Mixture	55.2	61.7	69.8	78.0
	SRU-NoSkips	66.3	76.7	84.5	90.5
	SRU-WithSkips	67.2	77.5	84.4	89.5
	SRU-AttenSkips	68.2	78.8	86.8	91.7

TABLE-V. PESQ in Seen Noise for IRM Training-Objective

Noise	Model	-8dB	-4dB	0dB	4dB
Babble Noise	Noisy Mixture	1.17	1.52	1.86	2.10
	SRU-NoSkips	1.65	1.88	2.17	2.51
	SRU-WithSkips	1.68	1.90	2.20	2.54
	SRU-AttenSkips	1.79	2.06	2.25	2.65
Car Noise	Noisy Mixture	1.27	1.42	1.62	1.87
	SRU-NoSkips	1.97	2.26	2.57	2.84
	SRU-WithSkips	2.01	2.29	2.60	2.87
	SRU-AttenSkips	2.10	2.34	2.66	2.96
Factory Noise	Noisy Mixture	1.28	1.32	1.52	1.76
	SRU-NoSkips	1.52	1.86	2.16	2.53
	SRU-WithSkips	1.55	1.87	2.17	2.55
	SRU-AttenSkips	1.62	2.01	2.26	2.67

TABLE-VI. PESQ in Seen Noise for IBM Training-Objective

Noise	Model	-8dB	-4dB	0dB	4dB
Babble Noise	Noisy Mixture	1.17	1.52	1.86	2.10
	SRU-NoSkips	1.63	1.85	2.17	2.50
	SRU-WithSkips	1.66	1.87	2.20	2.53
	SRU-AttenSkips	1.69	1.91	2.23	2.56
Car Noise	Noisy Mixture	1.27	1.42	1.62	1.87
	SRU-NoSkips	2.01	2.24	2.57	2.83
	SRU-WithSkips	2.04	2.27	2.61	2.86
	SRU-AttenSkips	2.07	2.30	2.63	2.89
Factory Noise	Noisy Mixture	1.28	1.32	1.52	1.76
	SRU-NoSkips	1.49	1.83	2.14	2.51
	SRU-WithSkips	1.50	1.84	2.15	2.52
	SRU-AttenSkips	1.59	1.92	2.23	2.59

Table V and Table VI assess the performance of the proposed SE models across seen noise types. According to PESQ evaluations, our SRU model, incorporating combined features and attention skips, demonstrated superior performance compared to models lacking skips or utilizing skips without an attention gate. This resulted in enhanced perceptual speech quality relative to counterpart models when applied to both noisy and proposed model-processed speech. For example in Table V and Table VI, SRU-AttSkip-IRM and SRU-AttSkip-IBM improved the PESQ by 0.34 (20.98%) and 0.31 (19.49%) over the noisy

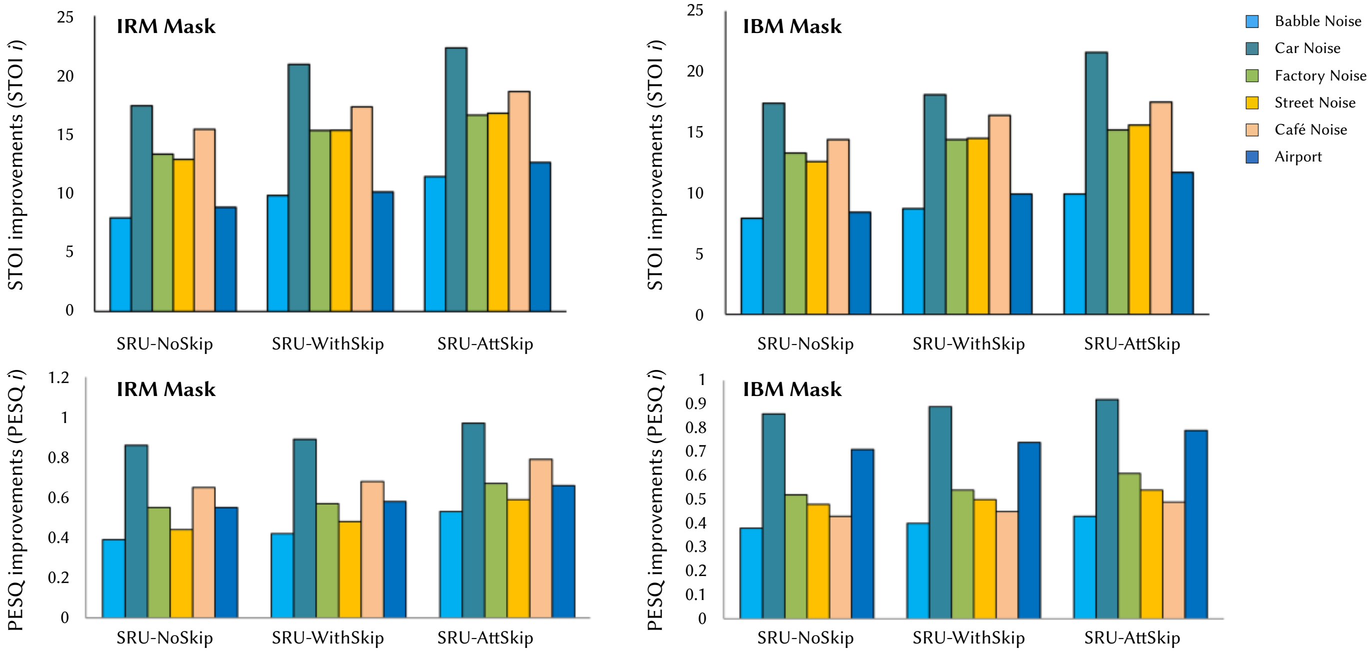


Fig. 6. Average enhancements (STOI*i* and PESQ*i*) across various noises.

TABLE-VII. STOI and PESQ Test Scores in All Example Noise Sources for TIMIT Dataset

Metric	Model	IRM				IBM			
		-8dB	-4dB	0dB	4dB	-8dB	-4dB	0dB	4dB
STOI	SRU-NoSkips	63.8	74.2	82.1	88.2	62.1	71.4	80.0	85.9
	SRU-WithSkips	65.5	78.2	84.2	90.4	63.8	75.5	82.1	88.0
	SRU-AttenSkips	67.3	79.5	85.6	91.8	66.2	77.6	84.8	90.7
PESQ	SRU-NoSkips	1.71	2.00	2.30	2.63	1.71	1.97	2.29	2.61
	SRU-WithSkips	1.75	2.02	2.32	2.65	1.73	1.99	2.32	2.64
	SRU-AttenSkips	1.84	2.14	2.39	2.76	1.78	2.05	2.36	2.68

mixture at -8dB factory noise. SRU-AttSkip-IRM and SRU-AttSkip-IBM improved the PESQ by 0.54 (26.21%) and 0.39 (20.41%) over noisy mixture at -4dB babble noise. Moreover, at 0dB car noise, the SRU-AttSkip-IRM and SRU-AttSkip-IBM enhanced PESQ by 1.04 (39.1%) and 1.01 (38.4%) over noisy mixture. Compare to other models, the proposed SRU-WithSkip-IRM and SRU-WithSkip-IBM with attention skips improved PESQ by 3.04% and 1.04% at 4dB in car noisy background. It shows that at high SNRs all the proposed SRU models perform nearly equally. Furthermore, models with attention skips improved the PESQ by 6.17% over SRU-NoSkip-IRM and 0.06 (2.34%) SRU-NoSkip-IRM at 4dB babble noise. For PESQ, SRU-AttSkip-IRM outscored SRU-AttSkip-IBM by 3.07%. Average STOI and PESQ for both masks can be found in Table VII and Table VIII, respectively. These scores are averaged across different seen noises. The obtained results validate that the proposed SRU-AttSkip achieved noteworthy results.

Average enhancements (STOI*i* and PESQ*i*) across various noises are illustrated in Fig. 6. Additional experimentation involved evaluating our SE models on the LibriSpeech dataset. This dataset, comprising 1000 hours of audiobook-derived utterances at 16 kHz sampling frequency, was chosen for evaluation. For this study, clean utterances were exclusively selected and mixed with noises (car, babble, airport, street, cafeteria). Table VIII shows PESQ and STOI for the LibriSpeech database. SRU-AttSkip-IRM and SRU-AttSkip-IBM configurations exhibited a significant 16.44% and 14.9% average STOI improvement over noisy speech. Correspondingly, these configurations led to an average PESQ improvement of 33.19% (0.78 factor) and 31.14% (0.71 factor) over the unprocessed speech. Cross-corpus comparisons highlighted the superior performance of the proposed models when trained on the LibriSpeech dataset in contrast to the TIMIT dataset.

Table IX presents the outcomes of the testing, focusing on CBAK and CSIG. The results clearly demonstrate that the robust feature sets and attention skips yielded superior performance in terms of both residual noise and distortion. In comparison, SRU-AttSkip-IRM and SRU-AttSkip-IBM effectively mitigated background noises and introduced less distortions when compared with SRU-NoSkip-IRM and SRU-NoSkip-IBM. Average C_{SIG} and C_{BAK} scores showed an enhancement from 2.02 and 1.73 with the noisy speech to 3.04, 3.01, 3.10, and 2.45 with SRU-NoSkip-IRM and SRU-NoSkip-IBM, marking progresses of 1.02 (33.55%) and 0.72 (29.4%), respectively. Correspondingly, C_{SIG} and C_{BAK} scores progressed from 3.04 and 2.45 with the SRU-NoSkip-IRM model to 3.10 and 2.49 with the SRU-WithSkip-IRM model. Lastly, C_{SIG} and C_{BAK} advanced from 3.05 and 2.43 with the SRU-WithSkip-IBM model to 3.13 and 2.48 with SRU-AttSkip-IBM.

Table X presents the performance in the seen noisy backgrounds. To evaluate our SE for noise generalization, Table XI presents the outcomes of PESQ and STOI tests for two unseen noises (cafeteria and factory2). Our SE models demonstrated significant performance over

TABLE-VIII. STOI and PESQ Test Scores in Five Example Noise Sources for LibriSpeech Dataset.

Metric	Model	IRM				IBM			
		-8dB	-4dB	0dB	4dB	-8dB	-4dB	0dB	4dB
STOI	SRU-NoSkips	63.9	75.0	81.7	88.3	62.9	71.1	80.4	86.3
	SRU-WithSkips	66.1	78.6	84.3	90.2	64.4	74.7	81.4	88.2
	SRU-AttenSkips	67.4	79.7	86.0	92.0	66.4	77.0	84.9	90.6
PESQ	SRU-NoSkips	1.72	2.04	2.30	2.66	1.66	1.98	2.28	2.61
	SRU-WithSkips	1.75	2.11	2.41	2.74	1.71	2.03	2.31	2.68
	SRU-AttenSkips	1.83	2.21	2.49	2.86	1.80	2.11	2.43	2.77

TABLE IX. C_{SIG} and C_{BAK} Test Scores in All Noise Sources at Four SNRs

Metric	C_{SIG}					C_{BAK}				
SNR (in dB)	-8	-4	0	4	Avg	-8	-4	0	4	Avg
Noisy Mixture	1.38	1.78	2.22	2.69	2.02	1.35	1.59	1.83	2.14	1.73
SRU-NoSkips-IRM	2.35	2.85	3.22	3.74	3.04	1.95	2.27	2.65	2.92	2.45
SRU-NoSkips-IBM	2.33	2.82	3.19	3.71	3.01	1.90	2.21	2.60	2.86	2.39
SRU-WithSkips-IRM	2.40	2.89	3.31	3.81	3.10	1.97	2.32	2.68	2.98	2.49
SRU-WithSkips-IBM	2.35	2.84	3.23	3.76	3.05	1.94	2.26	2.63	2.90	2.43
SRU-AttenSkips-IRM	2.54	3.01	3.41	3.91	3.22	2.06	2.40	2.78	3.06	2.58
SRU-AttenSkips-IBM	2.47	2.96	3.28	3.82	3.13	1.98	2.32	2.66	2.97	2.48

TABLE X. STOI and PESQ Test Scores in Seen Noise Sources Against Competing SE Algorithms

Metric	STOI					PESQ				
SNR (in dB)	-8	-4	0	4	Avg	-8	-4	0	4	Avg
Noisy Mixture	51.7	59.6	68.5	77.1	64.2	1.24	1.48	1.67	1.91	1.58
SRU-AttenSkips-IRM	65.5	76.0	83.9	90.3	79.0	1.81	2.09	2.39	2.72	2.25
SRU-AttenSkips-IBM	64.6	74.9	83.2	89.1	78.0	1.77	2.03	2.35	2.67	2.21
LSTM-IRM [3]	63.5	74.2	82.4	88.9	77.3	1.71	2.00	2.33	2.67	2.18
LSTM-IBM [3]	62.7	72.1	81.7	87.8	76.1	1.67	1.86	2.29	2.63	2.11
DNN-IRM [43]	58.5	70.0	78.7	85.6	73.2	1.57	1.75	2.19	2.53	2.01
DNN-IBM [43]	56.1	67.3	76.5	83.1	70.8	1.49	1.70	2.11	2.45	1.94
CNN [14]	59.3	70.0	79.8	86.8	74.0	1.62	1.83	2.25	2.59	2.07
GAN [3]	54.3	65.0	75.7	82.6	70.0	1.53	1.72	2.15	2.44	1.96

TABLE XI. STOI and PESQ Test Scores in Unseen Noise Sources Against Competing SE Algorithms

Metric	STOI					PESQ				
SNR (in dB)	-8	-4	0	4	Avg	-8	-4	0	4	Avg
Noisy Mixture	50.3	58.3	67.5	76.3	63.1	1.15	1.39	1.58	1.88	1.50
SRU-AttenSkips-IRM	64.3	74.8	82.7	90.0	78.0	1.79	2.05	2.34	2.69	2.22
SRU-AttenSkips-IBM	63.4	72.7	82.0	88.9	77.8	1.76	1.98	2.28	2.65	2.17
LSTM-IRM [3]	62.0	72.9	81.3	88.2	76.1	1.62	1.91	2.24	2.64	2.10
LSTM-IBM [3]	61.3	70.8	80.6	87.0	75.0	1.58	1.77	2.20	2.60	2.04
DNN-IRM [43]	57.0	68.6	77.7	84.8	72.0	1.48	1.66	2.10	2.51	1.94
DNN-IBM [43]	55.0	66.0	75.5	82.4	69.7	1.40	1.61	2.02	2.42	1.86
CNN [14]	57.9	68.7	78.8	86.0	72.9	1.53	1.74	2.16	2.56	2.00
GAN [3]	52.9	63.8	74.6	81.9	68.3	1.44	1.63	2.06	2.41	1.89

TABLE XII. STOI and PESQ Test Scores Against Unsupervised Competing SE Algorithms

Metric	STOI					PESQ				
SNR (in dB)	-8	-4	0	4	Avg	-8	-4	0	4	Avg
Noisy Mixture	50.3	58.3	67.5	76.3	63.1	1.15	1.39	1.58	1.88	1.50
SRU-AttenSkips-IRM	64.3	74.8	82.7	90.0	78.0	1.79	2.05	2.34	2.69	2.22
SRU-AttenSkips-IBM	63.4	72.7	82.0	88.9	77.8	1.76	1.98	2.28	2.65	2.17
LRSD [36]	54.3	63.2	70.6	79.2	66.8	1.38	1.71	1.98	2.28	1.83
NRPCA [33]	55.8	63.3	70.6	80.2	67.5	1.41	1.78	2.02	2.32	1.88
MMSE [6]	50.8	60.5	68.8	78.2	64.6	1.28	1.51	1.81	2.10	1.68

both baseline and competing networks in situations involving unseen noises. The most noteworthy STOI and PESQ scores were achieved by SRU-WithSkip-IRM and SRU-WithSkip-IBM models, owing to their sophisticated network architecture. With robust acoustic features and architectural changes applied to the proposed SRU models, the performance remained relatively stable across both seen and unseen noises. For instance, the average STOI values observed improvements of 14.9% and 13.7% over noisy speech when using SRU-WithSkip-IRM and SRU-WithSkips-IBM, respectively, resulting in STOI values of 78% and 76.8%. Notably, at SNRs (-4dB and -8dB) SRU-WithSkip-IRM and SRU-WithSkip-IBM exhibited STOI improvements of 1.89% and 1.78% as compared to baseline SRU (SRU-IRM and SRU-IBM). Furthermore, PESQ scores observed significant improvements, reaching 2.22 (32.43%) and 2.17 (31.90%) with SRU-WithSkip-IRM and SRU-WithSkip-IBM, respectively, compared to score of 1.50. This represents a substantial improvement in PESQ in unseen noisy conditions, outperforming the noisy speech. Across various metrics, our SRU models showcased considerable performance improvements compared to baseline SRU (SRU-IRM and SRU-IBM) and related models. Specifically, our SRU models exhibited STOI enhancements of 1.80%, 2.90%, 5.90%, 8.20%, 5%, and 9.6%, respectively, while also enhancing PESQ by factors of 0.10 (4.54%), 0.16 (7.27%), 0.26 (11.8%), 0.34 (15.45%), 0.20 (9.1%), and 0.31 (14.1%).

We present the results of the proposed models and their competitive counterparts, evaluated using STOI and PESQ metrics. The findings indicate notable improvements in speech quality, intelligibility, noise suppression, and speech distortions attributable to the proposed SRU. These models also showed superior performance when compared to the baseline SRU [25], DNN [12], CNN [43], and GAN (employing a 3-layer ReLU MLP) [19]. The results encompassing the proposed and competing models are tabulated in Table X and Table XI. The obtained results are averaged over all SNRs. The obtained scores show improvements in intelligibility and quality of the proposed SRU models for SE. For STOI, SRU-AttSkip-IRM and SRU-AttSkip-IBM surpassed DNN-IRM and DNN-IBM by 3.99% and 6.7%, respectively.

SRU-AttSkip-IRM showcased STOI improvement of 5.10% and 9.7% over CNN and GAN, whereas SRU-AttSkip-IBM showed an improvement of 8.49% over GAN and 5.01% over CNN. For PESQ, SRU-AttSkip-IRM, and SRU-AttSkip-IBM enhanced by 9.09% and 14.09% as compared to CNN and GAN, respectively. Moreover, a comparison with three unsupervised techniques was conducted to show success of supervised learning over unsupervised counterparts. These unsupervised methods include Low-rank sparse decomposition (LRSD) [44], Nonnegative RPCA (NRPCA) [45], and MMSE [46] for SE. Table XII provides the results. STOI scores showed improvements of 11.2%, 10.5%, and 13.4% with SRU-AttSkip-IRM, and 11%, 103%, and 13.2% with SRU-AttSkip-IBM. Similarly, PESQ scores showed improvements of 0.39 (equal to 17.56%), 0.34 (equivalent to 15.31%), and 0.54 (equal to 24.32%) with SRU-AttSkip compared to LRSD, NRPCA, and MMSE, respectively.

B. Spectrogram Analysis

To illustrate the spectral parts of speech that have undergone processing, this section presents a spectro-temporal analysis. Fig. 7 shows the spectrograms representing different speech utterances. In Fig. 7(a), we observe the spectrogram of clean speech. Fig. 7(b) shows that a clean sentence is mixed at 0dB with babble noise, resulting in noisy speech. This specific noisy condition is noteworthy because the noise characteristics resemble those of the target speech. Fig. 7(c) shows the enhanced speech through SRU-IBM, where background noise is evident. Figure 7(d) portrays the signals of SRU-IRM-enhanced speech, showcasing even lower levels of residual noise and distortion when compared to SRU-IBM. Figure 7(e) shows a spectrogram of the enhanced speech generated by SRU-AttSkip-IRM. This presentation indicates reduced distortion and residual noise. Concluding this experiment, Fig. 7(f) illustrates the enhanced speech attributed to SRU-AttSkip-IBM, revealing less distortions and residual noise.

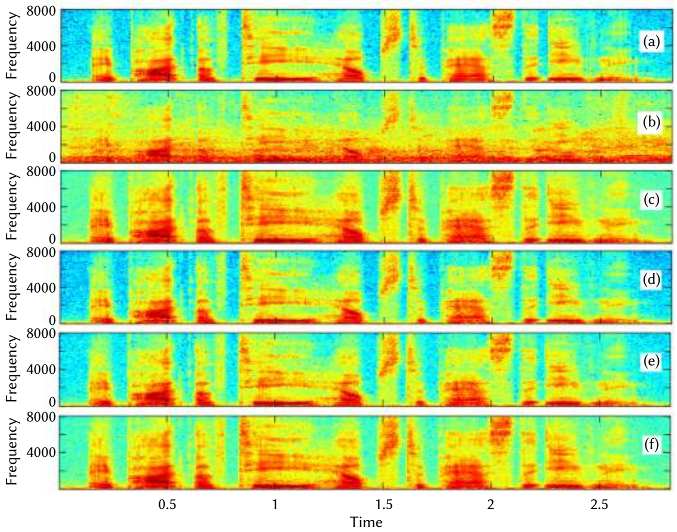


Fig. 7. Spectro-Temporal Analysis. (a) Clean, (b) Noisy, (c) SRU-IBM, (d) SRU-IRM, (e) SRU-AttSkip-IBM, and (f) SRU-AttSkip-IRM.

C. Computational Complexity

In practical applications, computational resources are frequently limited. Therefore, it is essential to strike a suitable balance between model performance and computational efficiency. Table XIII presents efficiency of parameters in the proposed SRU model. The assessment shows that implementing SRU-based SE with attention gates in skip connections considerably reduces the trainable parameters (2.607M) and parameter size (12.54MB), in comparison to similar models such as LSTM (17.384M, 65.42MB), GRU (13.33M, 52.33MB), and the baseline SRU (8.69M, 35.98MB). The introduction of attention gates into the skip connections leads to a slight increase in parameter count. In current DNN research for SE, there is a focus on enhancing model performance for hardware accelerators to execute these models rapidly and efficiently. To make the suggested SRU applicable in embedded systems, there is a need to minimize hardware memory consumption. This necessitates an examination of multiply–accumulate operations (MACs), where it is evident that the proposed SRU exhibits the lowest MACs (0.986 G/s with attention gates), ensuring efficient execution without compromising speech enhancement performance. Our SRU-based approach not only significantly reduces parameter size but also minimizes MACs. Additionally, the study evaluates the Real-Time Factor (RTF) in the proposed model, measuring the ratio of processing time to input audio data duration, which is crucial for real-time applications. Conducted on single-core Intel(R) Core (TM) i5-1135G7 CPU @ 2.40GHz processor, the experiment yields an RTF of 0.36.

TABLE XIII. Computational Complexity and Efficiency

Model	Para#	MACs	Memory
LSTM	17.38M	3.291 G/s	65.42 MB
GRU	13.33M	2.605 G/s	52.33 MB
SRU-(Baseline)	08.69M	1.554 G/s	35.98 MB
SRU-(Proposed)	2.607M	0.986 G/s	12.54 MB

D. Automatic Speech Recognition (ASR)

The results of the SE assessments demonstrate notable reduction of background noises and successfully recover the speech with a better intelligibility. Consequently, this study expects improved speech recognition capabilities, especially in presence of difficult noisy environments. Our speech enhancement model can be served as a front end preprocessor for obtaining lower word error rates (WER as ASR results). For ASR, this study implements the Kaldi toolkit [46], which adopts a GMM-HMM system and subsequently trains a DNN utilizing Mel-frequency filter-bank features. The training methodology draws inspiration from Tachioka [47]-[48]. Assessment of ASR performance is based on word error rates. For training the proposed SRU-driven speech enhancement models, a random selection of 2000 utterances was made from TIMIT and LibriSpeech. Following training the SRU models, the speech enhancement process was performed, leading to the synthesis of time-domain utterances. The synthesized time-domain utterances were used to train ASR models. As shown in Table XIV, ASR system presented improved performance when trained with data processed by SRU-AttSkip. The WERs showed a gradual reduction with more favorable SNR levels. On average, a WER of 14.25% was attained with the utilization of utterances processed by the proposed SRU-AttSkip. The results show that the proposed approach can effectively be utilized as a pre-processor to enhance ASR performance.

E. Performance Comparison Using VoiceBank+DEMAND

This section conducts experiments on the VoiceBank+DEMAND. The purpose of these experiments is to highlight the validation of the proposed SE approaches in comparison to contemporary benchmarks. The results of these experiments are detailed in Table XV. The analysis of results, as shown by PESQ, STOI, and Segmental SNR (SNRSeg), reveals certain observations. Higher values in these metrics signify enhanced performance. Notably, the experiments showcase results for SRU with skips, emphasizing a superior performance of attention skips with SRU. Our SRU perform better on VoiceBank+DEMAND and obtains competitive results: PESQ, STOI, SNRSeg, and trainable parameters. From GAGNet [49], the proposed SRU improves the metrics by 0.22 (PESQ), 0.5% STOI, and 0.64dB (SNRSeg), respectively. Further, from DCCRN [51], the proposed SRU improves the metrics by 0.47 (PESQ), 1.5% STOI, and 1.26dB (SNRSeg). The parameter count of the proposed SRU is better than all models except TSTNN [54]; however, PESQ = 2.96, STOI = 95.1%, and SNRSeg = 9.72dB are not reasonable to SRU. Our SRU obtains superior STOI, PESQ, and SNRSeg results with fewer trainable parameters (2.607M), MACs (0.986 G/s), and memory size (12.54MB).

F. Additional Experiments

We further examined the proposed SRU for different languages additional to English such as Urdu, Turkish, and Spanish. Although the speech enhancement aims to reduce the background noises from target speech, the results in Table XVI show that they are not severely degraded by different languages. However, since Urdu, Turkish, and Spanish are low-resource languages, the ASR performance is severely degraded in noisy environments. The STOI and PESQ values for all languages are marginally different which indicates that SE performance is not significantly affected by languages. To improve WERs, different strategies, such as speech augmentation and neural speech synthesis, can be used for ASR.

TABLE XIV. Computational Complexity and Efficiency

DNN Model	Noisy Mixture	SRU-AttSkip-IRM	SRU-AttSkip-IBM	SRU-IRM	SRU-IBM	DNN-IRM	DNN-IBM
WERs	55.35%	14.25%	14.75%	19.20%	19.95%	29.25%	30.02%

TABLE XV. Performance Evaluation on the VoiceBank+DEMAND Database

Model	Domain	Year	Parameter #	PESQ	STOI	SNRSeg
SEGAN [50]	Time	2017	97.5M	2.16	93.1	7.66
DCCRN [51]	Time-Frequency	2019	3.70M	2.68	93.7	8.62
GAGNet [49]	Time-Frequency	2021	5.64M	2.94	94.7	9.24
RDL-Net [52]	Time-Frequency	2020	3.91M	3.02	93.8	--
DEMUCS [53]	Time	2020	128M	3.07	95.1	8.53
TSTNN [54]	Time	2021	0.92M	2.96	95.1	9.72
SE-Conformer [55]	Time-Frequency	2021	--	3.13	95.1	--
PFR-Net [58]	Time-Frequency	2022	4.61M	3.19	95.0	--
FAF-Net [59]	Time-Frequency	2022	6.90M	3.24	95.0	--
MAB-CED [60]	Time-Frequency	2022	4.82M	2.84	85.0	--
SRU (Baseline)	Time-Frequency	2024	8.69M	3.09	94.4	9.11
SRU (Proposed)	Time-Frequency	2024	2.61M	3.15	95.4	9.88

TABLE XVI. ASR for Different Languages

Model	SRU-AttSkip-IRM			SRU-AttSkip-IBM		
Language	STOI	PESQ	WERs	STOI	PESQ	WERs
English	78.12	2.19	14.25	78.03	2.15	14.75
Urdu	78.10	2.14	30.14	77.94	2.10	31.21
Turkish	77.95	2.09	20.22	77.84	2.04	22.14
Spanish	77.47	2.01	26.11	77.34	1.98	27.25

VI. Conclusions

In this study, a novel speech enhancement (SE) system is introduced, utilizing lightweight recurrent neural networks (RNNs) that have been trained with robust features. The approach includes the development of an hourglass SRU model, which effectively captures temporal dependencies by decreasing feature resolutions. To counteract gradient decline across layers, nonadjacent symmetrical layers are connected through skip connections. Furthermore, an attention gate is integrated into these skips, aiming to emphasize significant features and spectral regions. Composite and robust features are derived from the noisy magnitude, enhancing training of the proposed models for improved SE and ASR performance. The model independently estimates two masks: ideal ratio mask and ideal binary mask. The findings show several key aspects of the proposed speech enhancement model. The incorporation of combined feature learning allows for the integration of additional information, enabling the network to grasp the intricate nonlinear relationship between noisy and clean speech. The adopted SRU architecture proficiently captures long-term temporal dependencies while downsizing the feature resolution for parameter estimation during testing, thereby preventing excessive memory usage. The introduction of the skips and attention gates within these skips significantly addressed gradient decay across layers, additionally highlighting important features and spectral regions. The proposed SRU strategy contributes to superior performance compared to the baseline, as evident from enhanced trainable parameter metrics. The proposed speech enhancement model outperforms the recent deep SE representations across diverse background noises, showing promising outcomes concerning speech distortion and residual noise. Notably, the models showcase superior results not only in seen noisy environments but also in unseen noise contexts. Results from GMM-HMM ASR show the potential of SRU-AttSkip SE model as a preprocessor for enhancing ASR performance in noisy environments. Our SRU model performs better on the VoiceBank+DEMAND and obtains competitive results: PESQ, STOI, SNRSeg, and trainable parameters. Further, the proposed SRU has the lowest MACs (0.986 G/s) with attention gates. SRU obtains superior STOI, PESQ, and SNRSeg results with fewer trainable parameters (2.607M), MACs (0.986 G/s), and memory size (12.54MB).

The SRU model can be sensitive to the noisy inputs. If the input data contains significant noise or errors, the model's performance may degrade. In the future, we anticipate the extension of the proposed network architecture to address this limitation and use for regression-based speech enhancement (SE), jointly optimized with application to automatic speech recognition and automatic speaker recognition. Additionally, the potential for devising more robust acoustic features is observed, offering the prospect for further work.

Acknowledgment

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through large group Research Project under grant number RGP2/383/44.

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R51), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Funding

The current work was supported by the Deanship of Scientific Research at King Khalid University through large group Research Project under grant number RGP2/383/44.

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R51), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113– 120, 1979.

[2] S. Nasir, A. Sher, K. Usman, U. Farman, "Speech enhancement with geometric advent of spectral subtraction using connected time-frequency regions noise estimation", *Research Journal of Applied Sciences,* Engineering and Technology, vol. 6, no. 6, pp. 1081–1087, 2013.

[3] J. Lim, A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[4] Y. Ephraim, D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[5] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[6] N. Mohammadiha, P. Smaragdis, A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 21, no. 10, pp. 2140–2151, 2013.

[7] I. Tashev, M. Slaney, "Data driven suppression rule for speech enhancement," in: *2013 Information Theory and Applications Workshop (ITA)*, IEEE, 2013, pp. 1–6.

[8] Y. Xu, J. Du, L.-R. Dai, C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2013.

[9] Y. Xu, J. Du, L.-R. Dai, C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[10] M. Kolbæk, Z.-H. Tan, J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 25, no. 1, pp. 153–167, 2016.

[11] S. Hochreiter, J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] Y. Wang, A. Narayanan, D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[13] N. Saleem, M.I. Khattak, "Deep neural networks for speech enhancement in complex-noisy environments", *International Journal of Interactive Multimedia and Artificial Intelligence,* vol. 6, no. 1, pp. 84–91, 2020.

[14] N. Saleem, M.I. Khattak, M. Al-Hasan, A.B. Qazi, "On learning spectral masking for single channel speech enhancement using feedforward and recurrent neural networks," *IEEE Access,* vol. 8, pp. 160581–160595, 2020.

[15] N. Saleem, M.I. Khattak, "Multi-scale decomposition based supervised single channel deep speech enhancement," *Applied Soft Computing*, vol. 95, pp. 106666, 2020.

[16] N. Saleem, M.I. Khattak, M. Al-Hasan, A. Jan, "Multi-objective long-short term memory recurrent neural networks for speech enhancement," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 10, pp. 9037– 9052, 2021.

[17] S. Samui, I. Chakrabarti, S.K. Ghosh, "Time–frequency masking based supervised speech enhancement framework using fuzzy deep belief network," *Applied Soft Computing*, vol. 74, pp. 583–602, 2019.

[18] M.H. Soni, N. Shah, H.A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 5039–5043.

[19] N. Shah, H.A. Patil, M.H. Soni, "Time-frequency mask-based speech enhancement using convolutional generative adversarial network," in: *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2018, pp. 1246–1251.

[20] W. Yu, J. Zhou, H. Wang, L. Tao, "Setransformer: speech enhancement transformer," *Cognitive Computation,* vol. 14, pp. 1152-1158, 2022.

[21] J. Cadore, F.J. Valverde-Albacete, A. Gallardo-Antolín, C. Peláez-Moreno, "Auditory-inspired morphological processing of speech spectrograms: Applications in automatic speech recognition and speech enhancement," *Cognitive computation,* vol. 5, pp. 426–441, 2013.

[22] I. Sutskever, O. Vinyals, Q.V. Le, "Sequence to sequence learning with neural networks," in *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, vol. 2, 2014, pp. 3104–3112.

[23] I. Serban, A. Sordoni, Y. Bengio, A. Courville, J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[24] K. Zarzycki, M. Ławryńczuk, "LSTM and GRU neural networks as models of dynamical processes used in predictive control: A comparison of models developed for two chemical reactors," *Sensors*, vol. 21, no. 16, pp. 5625, 2021.

[25] J. Chen, D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America,* vol. 141, no. 6, pp. 4705–4714, 2017.

[26] M. Sundermeyer, H. Ney, R. Schlüter, "From feedforward to recurrent lstm neural networks for language modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 23, no. 3, pp. 517–529, 2015.

[27] M. Fernández-Díaz, A. Gallardo-Antolín, "An attention long short-term memory based system for automatic classification of speech intelligibility," *Engineering Applications of Artificial Intelligence,* vol. 96, pp. 103976, 2020.

[28] M. Q. Gandapur, E. Verdú, "ConvGRU-CNN: Spatiotemporal Deep Learning for Real-World Anomaly Detection in Video Surveillance System," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 8, no. 4, 2023.

[29] N. Saleem, J. Gao, M.I. Khattak, H.T. Rauf, S. Kadry, M. Shafi, "Deepresgru: Residual gated recurrent neural network-augmented kalman filtering for speech enhancement and recognition," *Knowledge-Based Systems,* vol. 238, pp. 107914, 2022.

[30] J. Ali Reshi, R. Ali, "An Efficient Fake News Detection System Using Contextualized Embeddings and Recurrent Neural Network," *International Journal of Interactive Multimedia and Artificial Intelligence*, pp. 1-13, 10.9781/ijimai.2023.02.007.

[31] B. Chang, L. Meng, E. Haber, F. Tung, D. Begert, "Multi-level residual networks from dynamical systems view," arXiv preprint arXiv:1710.10348, 2017.

[32] Y. Shao, S. Srinivasan, Z. Jin, D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Computer Speech & Language,* vol. 24, no. 1, pp. 77–93, 2010.

[33] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom," NIST speech disc 1-1.1. NASA STI/Recon technical report, no. 93, 27403, 1993.

[34] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[35] D. Pearce, J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," *Inst. for Signal & Inform. Process.*, Mississippi State Univ., Tech. Rep, 2002,

[36] A. Varga, H.J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247– 251, 1993.

[37] T. F. Damayanti, A. Wanto, H.S. Tambunan, "Prediction of Palm Oil Seed Stock Production Results with the Back-propagation Algorithm," *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, vol. 2, no. 2, pp. 105-112, 2023.

[38] Q. Song, Y. Wu, Y.C. Soh, "Robust adaptive gradient-descent training algorithm for recurrent neural networks in discrete time domain," *IEEE Transactions on Neural networks,* vol. 19, no. 11, pp. 1841–1853, 2008.

[39] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv preprint arXiv:1207.0580, 2012.

[40] A.W. Rix, M.P. Hollier, A.P. Hekstra, J.G. Beerends, "Perceptual evaluation of speech quality (PESQ) the new itu standard for end-to-end speech quality assessment part i–time-delay compensation," *Journal of the Audio Engineering Society,* vol. 50, no. 10, pp. 755–764, 2002.

[41] C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2010, pp. 4214–4217.

[42] Y. Hu, P.C. Loizou, "Evaluation of objective measures for speech enhancement," in: *Ninth International Conference on Spoken Language Processing*, 2006.

[43] T. Kounovsky, J. Malek, "Single channel speech enhancement using convolutional neural network," in: *2017 IEEE International Workshop of Electronics, Control, Measurement, Signals and Their Application to Mechatronics (ECMSM)*, IEEE, 2017, pp. 1–5.

[44] P. Sun, J. Qin, "Low-rank and sparsity analysis applied to speech enhancement via online estimated dictionary," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1862–1866, 2016.

[45] W. Shi, X. Zhang, X. Zou, W. Han, G. Min, "Auditory mask estimation by RPCA for monaural speech enhancement," in: *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, IEEE, 2017, pp. 179–184.

[46] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., "The kaldi speech recognition toolkit," in: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society, 2011.

[47] Y. Tachioka, S. Watanabe, J. Le Roux, J.R. Hershey, "Discriminative methods for noise robust speech recognition: A chime challenge benchmark," in: *The 2nd International Workshop on Machine Listening in Multisource Environments*, 2013, pp. 19–24.

[48] A. Shewalkar, D. Nyavanandi, S.A. Ludwig, "Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU," *Journal of Artificial Intelligence and Soft Computing Research,* vol. 9, no. 4, pp. 235-245, 2019.

[49] A. Li, C. Zheng, L. Zhang, X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement" *Applied Acoustics*, vol. 187, 108499, 2022.

[50] S. Pascual, J. Serra, A. Bonafonte, "Time-domain speech enhancement using generative adversarial networks," *Speech communication*, vol. 114, pp. 10-21, 2019.

[51] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, ... & L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," arXiv preprint arXiv:2008.00264, 2020.

[52] M. Nikzad, A. Nicolson, Y. Gao, J. Zhou, K.K. Paliwal, F. Shang, "Deep residual-dense lattice network for speech enhancement," in: *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 34, no. 05, 2020, pp. 8552-8559.

[53] A. Defossez, G. Synnaeve, Y. Adi, "Real time speech enhancement in the waveform domain," arXiv preprint arXiv:2006.12847, 2020.

[54] K. Wang, B. He, W.P. Zhu, "TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 7098-7102.

[55] E. Kim, H. Seo, "SE-Conformer: Time-Domain Speech Enhancement Using Conformer," in *Interspeech,* 2021, pp. 2736-2740.

[56] Z. Ye, N. Saleem, H. Ali, "Efficient Gated Convolutional Recurrent Neural Networks for Real-Time Speech Enhancement," *International Journal of Interactive Multimedia and Artificial Intelligence*, 2023, doi: 10.9781/ijimai.2023.05.007.

[57] M.I. Khattak, A. Jan, N. Saleem, E. Verdú, N. Khurshid, "Automated detection of COVID-19 using chest X-ray images and CT scans through multilayer-spatial convolutional neural networks," *International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6*, no. 6, pp. 15-24, 2021.

[58] G. Yao, C. Wang, Y. Wu, Y. Wang, "Pyramid fully residual network for single image de-raining," *Neurocomputing*, vol. *456*, pp.168-178, 2021.

[59] H. Yue, W. Duo, X. Peng, J. Yang, "Reference-based speech enhancement via feature alignment and fusion network," in: *Proceedings of the AAAI*

Conference on Artificial Intelligence, vol. 36, no. 10, 2022, pp. 11648-11656.

[60] N. Saleem, T.S. Gunawan, M. Shafi, S. Bourouis, A. Trigui, "Multi-Attention Bottleneck for Gated Convolutional Encoder-Decoder-Based Speech Enhancement," *IEEE Access*, vol. *11*, pp. 114172-114186, 2023

Sami Dhahbi

Sami Dhahbi received the engineering and M.S. degrees from the National School of Computer Science, University of Manouba, Tunisia, in 2005 and 2006, respectively, and the Ph.D. degree in computer science from the University of Tunis, El Manar, Tunisia, in 2016. He is currently an assistant professor of computer science at King Khalid University, Saudi Arabia. He is also a member of the LIMTIC Research Laboratory at the University of Tunis, El Manar. He is the author of several articles. His research interests include machine learning, medical imaging, and more recently, networks and cloud computing.

Nasir Saleem

Nasir Saleem received B.S. M.S. and PhD Engineering degree from University of Engineering & Technology, Peshawar-25000, Pakistan, in 2008; 2012, and 2021 with specialization in speech processing and deep learning. Did postdoctoral fellow, Islamic International University Malaysia (IIUM), researching the artificial intelligence-based speech processing algorithms. From 2008 to 2012, he was a senior lecturer at the Institute of Engineering Technology (IET), Gomal University, where he was involved in teaching and research. He is now an assistant professor in the Department of Electrical Engineering, Faculty of Engineering and Technology (FET), and Deputy Director of the Quality Assurance Directorate at Gomal University. Human-machine interaction, speech enhancement, speech recognition, speech and video processing, and machine learning applications are the areas he is currently researching.

Teddy Surya Gunawan

Teddy Surya Gunawan (Senior Member, IEEE) received his BEng degree in Electrical Engineering with cum laude award from Institut Teknologi Bandung (ITB), Indonesia, in 1998. He obtained his M.Eng degree in 2001 from the School of Computer Engineering at Nanyang Technological University, Singapore, and a Ph.D. degree in 2007 from the School of Electrical Engineering and Telecommunications, The University of New South Wales, Australia. He was a Visiting Research Fellow (2010 to 2021) at UNSW and is currently an Adjunct Professor at Telkom University (2022-2023). His research interests are speech and audio processing, biomedical signal processing and instrumentation, image and video processing, and parallel computing. He was awarded the Best Researcher Award in 2018 from IIUM. He is currently an IEEE Senior Member (since 2012), was chairperson of the IEEE Instrumentation and Measurement Society – Malaysia Section (2014, 2020, 2021), Professor (since 2019), Head of Department (2015-2016) at the Department of Electrical and Computer Engineering, and Head of Programme Accreditation and Quality Assurance for Faculty of Engineering (2017-2018), International Islamic University Malaysia. He has been a Chartered Engineer (IET, UK) since 2016 and Insinyur Profesional Utama (PII, Indonesia) since 2021, a registered ASEAN engineer since 2018, and ASEAN Chartered Professional Engineer since 2020.

Sami Bourouis

Sami Bourouis received the Engineer, M.Sc., and Ph.D. degrees in computer science from the University of Tunis, Tunisia, in 2003, 2005, and 2011, respectively. He is currently a Professor at the College of Computers and Information Technology, Taif University, Saudi Arabia. His research interests include data mining, image processing, statistical machine learning, cybersecurity, and pattern recognition applied to several real-life applications.

Imad Ali

Imad Ali received B.Sc. Telecom Engineering degree from the University of Engineering and Technology, Peshawar, Pakistan in 2008, M.S. Electrical Engineering degree from CECOS University, Peshawar, Pakistan in 2012; and a Ph.D. degree in Social Networks and Human-Centered Computing, from National Tsing Hua University, Taiwan, in collaboration with Academia Sinica, Taiwan, in 2020. From 2009 to 2014, he served as a lecturer at different universities in Pakistan. He is currently working as an Assistant Professor in the Department of Computer Science, University of Swat, Pakistan. His research interest includes Question Answering Systems, Data Science, and Machine Learning.

Aymen Trigui

Aymen Trigui received the Engineer, M.Sc., and Ph.D. degrees in computer science from the University of Tunis, Tunisia, in 2003, 2005, and 2011, respectively. He is currently an Associate Professor at the Department of Computer Science, College of Computer Science, King Khalid University, Abha, Saudi Arabia. His research interests include data mining, image processing, statistical machine learning, cybersecurity, and pattern recognition applied to several real-life applications.

Abeer D. Algarni

Abeer D. Algarni received the B.Sc. degree (Hons.) in computer science from King Saud University, Riyadh, Saudi Arabia, in 2007, and the M.Sc. and Ph.D. degrees from the School of Engineering and Computer Sciences, Durham University, U.K., in 2010 and 2015, respectively. She has been an Assistant Professor with the College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, since 2008. Her current research interests include networking and communication systems, digital image processing, digital communications, and cyber security.

Drug Target Interaction Prediction Using Machine Learning Techniques – A Review

A. Suruliandi[1], T. Idhaya[1], S. P. Raja[2] *

[1] Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Abhishekapatti, Tirunelveli, TamilNadu (India)
[2] School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, TamilNadu (India)

* Corresponding author. suruliandi@yahoo.com (A. Suruliandi), idhayathomas003@gmail.com (T. Idhaya), avemariaraja@gmail.com (S. P. Raja).

Received 13 August 2021 | Accepted 4 January 2022 | Early Access 10 November 2022

Abstract

Drug discovery is a key process, given the rising and ubiquitous demand for medication to stay in good shape right through the course of one's life. Drugs are small molecules that inhibit or activate the function of a protein, offering patients a host of therapeutic benefits. Drug design is the inventive process of finding new medication, based on targets or proteins. Identifying new drugs is a process that involves time and money. This is where computer-aided drug design helps cut time and costs. Drug design needs drug targets that are a protein and a drug compound, with which the interaction between a drug and a target is established. Interaction, in this context, refers to the process of discovering protein binding sites, which are protein pockets that bind with drugs. Pockets are regions on a protein macromolecule that bind to drug molecules. Researchers have been at work trying to determine new Drug Target Interactions (DTI) that predict whether or not a given drug molecule will bind to a target. Machine learning (ML) techniques help establish the interaction between drugs and their targets, using computer-aided drug design. This paper aims to explore ML techniques better for DTI prediction and boost future research. Qualitative and quantitative analyses of ML techniques show that several have been applied to predict DTIs, employing a range of classifiers. Though DTI prediction improves with negative drug target pairs (DTP), the lack of true negative DTPs has led to the use a particular dataset of drugs and targets. Using dynamic DTPs improves DTI prediction. Little attention has so far been paid to developing a new classifier for DTI classification, and there is, unquestionably, a need for better ones.

Keywords

Chemogenomics, Drug Databases, Drug Discovery, Drug Target Interactions, Machine Learning, Targets, Target Databases.

DOI: 10.9781/ijimai.2022.11.002

I. Introduction

Discovering new drugs is critical and driven by the need for medication in daily life, partly brought on by changing environmental conditions. Nevertheless, drug discovery is not easy, it demands time as well as money, and the drug success rate is usually low. Computer-Aided Drug Design (CADD) is considered a computational discipline that aims to discover, design, and develop therapeutic chemical targets. There are 3 phases in drug design - discovery, development, and registry.

In the first phase, discovery, the focus is on identifying a new drug and its targets, based on binding sites. The second phase, development, involves pre-clinical research, where the drug is tested on animals for safety. Successful research means that human trials are set in motion. In the third phase, registry, the Food and Drug Administration (FDA) thoroughly reviews all the submitted drug-related data and decides on its approval or otherwise. Initiating an efficient computational model that finds potential Drug Target Interaction (DTI) from biological data helps understand the biological process, recognize novel drugs, and offer improved therapeutic medicine for illnesses of all sorts. Drug development has three trial phases, each of which is more expensive than the others. As of today, the cost of drug development has risen from US$3.4 million to US$8.6 million and US$21.4 million for phase I, phase II and phase III trials, respectively [1]. A new drug could fail to pass the test in any of the three drug development trial phases, notwithstanding the expense, effort and time involved.

II. State of the Art Methods

DTI is the process of finding new drugs and targets for drug development. Drug and target molecules are discovered through their interactions. Drug discovery methods are ligand-based, docking-based and chemogenomics-based, and involve parameters like biomarker identification, structure unavailability, physique and condition, and environmental factors. Current research is focused on maximizing interactions so the drugs formulated can successfully treat disease.

Please cite this article as:
A. Suruliandi, T. Idhaya, 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, http://dx.doi.org/10.9781/ijimai.2022.11.002

The new drugs developed today, though based on knowledge of existing ones, could still have adverse side effects. Incidentally, a drug developed for a particular disease may be used, quite unexpectedly, to treat another disease with no side effects whatsoever, a process referred to as drug repurposing [2], [3]. It is essential in drug discovery to establish the interaction between a drug and a target gene. The docking-based method needs a 3D structure of the target protein or gene for the process to work. The success of a newly developed drug depends on how well it fares in the market, particularly in terms of whether the purpose for which it was originally designed is being fulfilled. The possibility of successfully identifying DTI is enhanced by working on binding factors or interacting sites. This is a difficult process, given the limited information on drugs and targets. Bioinformaticians have tried to draw information from factors driving drugs and targets. The automated tools employed to improve the success rate by discovering more interactions or binding sites between drugs and their targets are intended to actively assist doctors and bioinformaticians. Scientists today work in drug development using ML predictive analysis techniques to understand drugs and targets, thus boosting DTI success prediction.

A. Drug Developing Procedure

Drugs are synthesized chemicals that control, prevent, and cure and diagnose illnesses. Disease diagnosis is carried out through reading the body's reactions to drug molecules in the form of positive biological responses. In pharmacological terms, the biomolecule whose function and activity are modified by a specific drug is termed the drug target. Biomolecules can be proteins, nucleic acids, receptors, enzymes, and ion channels. The DTI process interacts or binds the drug molecule to the active biomolecule site with the same structural or functional properties as the drug molecule, culminating in the creation of a new product as in Fig.1. The human body assimilates the product, resulting in a cure.

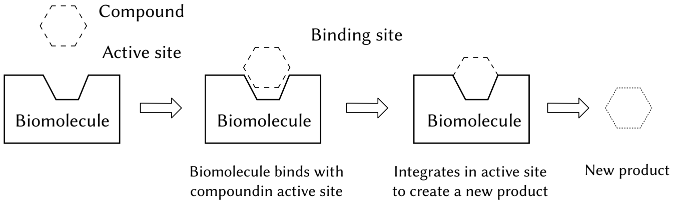

Compound
Active site
Binding site
Biomolecule
Biomolecule
Biomolecule
Biomolecule binds with compoundin active site
Integrates in active site to create a new product
New product

Fig.1. Drug Developing Procedure.

Drugs are developed in three phases. In the first phase, a drug and its target are discovered by means of the interacting or binding site, using substrate on the active site of protein. In the second phase, the drug is subjected to animal testing for safety's sake. In the third phase, the drug has human trials, following which it is marketed.

B. In-Silico Approaches in Drug Discovery

In-vitro is a technique where the process of drug discovery takes place in a controlled environment but not within a living organism. Here a pool of potential compounds is identified and narrowed down to find most reliable compound for treatment. In-vivo is a technique where the process of drug discovery takes place within a living organism by giving the reliable compounds to the human trials. Both the data collected from in-vitro and in-vivo are given as input features to the in-silico methods for drug prediction, which is a computational method. The computational DTI prediction method is categorized into the three approaches [4].

1. Docking-Based Approach

A docking-based approach in DTI prediction requires a 3D structure for simulation. Consequently, it is not applicable where a large number of proteins is involved, as in, for instance, the G-Couple Protein Receptor and ion channel, whose structures are far too complex to be obtained. The simulation is significant in regard to the time taken and its overall efficiency.

2. Ligand-Based Approach

A ligand-based approach works on the premise that a drug can be predicted without the 3-Dimensional structure of targets and with the existing knowledge of drugs and its targets.

3. Chemogenomics-Based Approach

A chemogenomics-based approach integrates both the chemical space of drugs and the genomic space of targets into a single pharmacological space. The challenge here is that there are too few DTI pairs and too many unknown interaction pairs.

C. Motivation and Justification

The in-vitro prediction of DTI from biological data calls for a lot of effort in the search for new drugs and targets. Identifying potential drugs and targets is a painstaking step in initiating drug discovery. Despite the plethora of research on DTI prediction in the recent past, prediction is still material-intensive and protracted. Predicting interaction between DTPs continues to challenge researchers. The motivation for this review is to help researchers in the drug development domain access state-of-the-art methods used in ML for DTI predictions, and so enhance the quality of research. To this end, several insightful articles on DTI procedures and methods that help discover new drugs and targets differently are reviewed. The machine learning (ML) techniques used to predict DTIs are studied, each with its strengths and limitations. The research is categorized, based on the ML techniques used in the prediction. Thereafter, it is qualitatively and quantitatively analyzed to understand ML and DTI better so the latter can be improved.

The contributions of this paper are as follows, Articles related to ML and DTI in drug development are studied in detail and categorized, based on the machine learning techniques deployed as in section III. The feature selection techniques used in DTI prediction suggest the best features for use. Articles on DTI prediction using ML techniques have described how ML manages datasets from miscellaneous databases, balances imbalanced data, handles large-scale datasets and features and, finally, examines at length the ML algorithms used in DTI prediction. Articles that are qualitatively analysed in section V based on ML techniques to understand their strengths and weaknesses. A quantitative analysis in section VI follows to find the most appropriate classifiers for DTI predictions.

D. Organization of the Paper

The paper is organized as follows. Section II provides an overview of state of the art methods involved in DTI prediction using ML techniques. In Section III, Machine learning techniques used for DTI prediction are summarized. In Section IV, databases used for DTI prediction are discussed. In Section V, a qualitative analysis of the ML techniques used for DTI is presented. In Section VI, a quantitative analysis of DTI prediction methods is offered. Section VII discusses DTI prediction. Section VIII concludes the study and offers new directions for future research.

III. Machine Learning (ML) Techniques Used for DTI Prediction

Computational models use ML techniques for prediction because they optimize data better and perform better as well. ML techniques, which learn data without relying on previously defined formulas, are grouped into two – supervised and unsupervised learning. Supervised learning predictions are based on observed existing knowledge from known data, while unsupervised learning predictions do the same without. Predictions are guesses based on existing knowledge from the data at hand. On the other hand, Classification refers to the process of differentiating between known and unknown labels. The objective of this paper is to explore ML techniques involved in improving DTP identification to find DTIs. The identification of a new drug involves the drug and its target. Because of large number of features of both drugs and targets manually extracting them would be a time taking process, so the researchers use only tools like ChemCPP, EDragon, CDK, Open Babel, RDkit, PADEL for extracting the features from drugs and Protr, SPICE, Propy, ProtDcal, ProtParam for extracting features from targets. Drug and target features are extracted and concatenated with each other to form DTPs. The pairs are analyzed for interaction prediction; specifically, to observe whether or not the DTPs interact. The ML techniques analyzed are explained qualitatively and quantitatively and the classifier used for DTI prediction is found. The DTI prediction here mainly uses a static database. Prediction can be improved when there are more targets and drugs with the interaction between them yet to be ascertained. In recent times, CADD has been used to develop drugs for immunodeficiency syndrome, influenza virus infection, glaucoma and lung cancer [5]. CADD helps in pharmacological, Pharmacodynamics and in-silico toxicity prediction, which identifies or filters inactive or toxic molecules [6] and naturally gets ML involved in DTI prediction strategies [7]-[10]. Thus to improve drug development various methods based on drugs and targets are developed using ML techniques. Fig.2 shows DTI prediction through ML techniques with targets and drugs taken from diverse databases. Drug and target features are extracted using a slew of tools or web servers. Subsequently, the most influential features alone are selected and used for DTI prediction with several ML classifiers to complete the process.

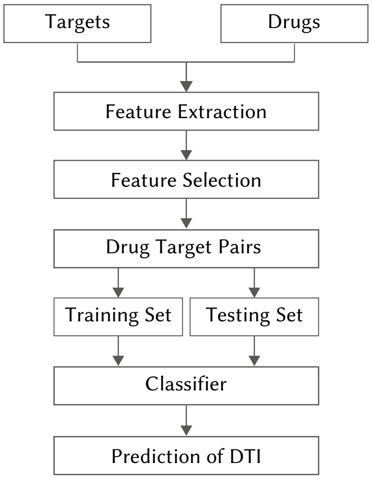


Fig.2. Flow of DTI Prediction.

In-silico methods include Machine learning, Data mining, Network analysis tool and data analysis tool, Quantitative Structure Analysis Relationship (QSAR), pharmacophores, homology modeling, Here Machine learning technique is more feasible than all other methods for working with drug discovery data for analysis. The trending research in drug discovery is "Identification of screening hits (compounds)" which helps in finding the particular compounds target with more potency at different level like binding, reducing the side effects, efficiency, and also increases the life of patients by changing the function of the biomolecule.

A. Chemogenomics-Based Machine Learning (ML) Techniques for DTI Prediction

The chemogenomics-based prediction approach is computationally predicted using ML-based, graph-based or network-based methods. ML-based methods are explained below in Fig 3.

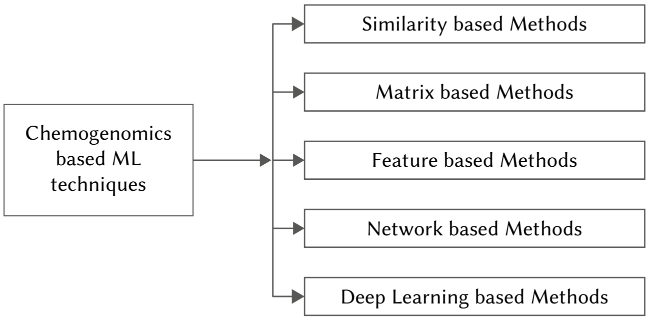


Fig.3. Chemogenomics based ML Techniques.

1. Similarity-Based Methods

The most commonly used DTI prediction methods use drug and target similarity measures in tandem with the distance between each pair of drugs and its targets [11]-[18]. These methods use the drug, target and drug-target interaction similarity scores based on prior knowledge of their interaction similarity. The similarity is obtained using a distance function like the Euclidean. For instance, if the following function is employed for the nearest neighbor algorithm, assuming two vectors x1 and x2, the distance between the vectors is found using equation (1) as D(x1, x2) where

$$D\ (x1, x2) = 1 - \frac{x1.x2}{||x1||||x2||} \tag{1}$$

and the same dimension and distance are calculated using the Euclidean norm and the inner product. The similarity between a drug and a target is given through the pharmacological similarity of the drug, the genomic similarity of the protein sequence, and the topological properties of a multipartite network of previously known drug-target interaction knowledge. The disadvantage of these methods is that they use knowledge drawn from a small quantum of labelled data, while there exist large quanta of unlabeled data.

2. Matrix-Based Methods

Several studies [19]-[24] have shown that matrix-based methods outperform the rest in DTI prediction. The interaction matrix is

$$X_{mxn} = \begin{bmatrix} x11 & \cdots & x1n \\ \vdots & \vdots & \vdots \\ xm1 & \cdots & xmn \end{bmatrix} \tag{2}$$

For i=1: m and j=1: n,

$$X_{ij} = \begin{cases} \text{1 if drug i and target j interact} \\ \text{0 there is no interaction between drug i and target j} \end{cases}$$

The first move in DTI prediction is to break down matrix X_{mxn} into two matrices, Y_{mxk} and Z_{nxk}, where X ~ YZ^T with k < m, n, and where Z^T denotes the swapped matrix of Z. This process of factorizing matrices in lower order makes it easier for matrix-based approach to deal with the missing data. With these methods, however, the distance between the drug and target appears to be the same and establishes the strength of the interaction between them, embedding them in a low-dimensional matrix. The reliability of these methods is affected when the drug and target data increase in volume, impacting the capacity to find their interaction.

3. Feature-Based Methods

Feature-based prediction methods largely use the support vector machine to find drug-target interaction [25]-[33]. Any pair of targets and drugs may be represented with features, leading to binary classification or two-class clustering with positive or negative interactions. Features are represented as F

$$F= \{d + t\},\ d= d_1,\ d_2,d_3,....d_a \text{ and } t = t_1,\ t_2,\ t_3,\ ....\ t_b \tag{3}$$

where d denotes the drug features of length a and t the target features of length b, respectively.

4. Network-Based Methods

Network-based methods [34]-[40], which use graph-based techniques to predict DTI, are considered simple and reliable interaction prediction methods. Here, the drug-drug similarity, target-target similarity and known interactions between DTI are integrated into a heterogeneous network, operating on the simple logical principle that similar drugs interact with similar targets.

5. Deep Learning-Based Methods

Deep learning-based approaches can reduce the loss of feature information in predicting DTIs. However, they need adequate information to predict interaction and drug repurposing [41]-[45]. The two steps of deep learning include generating feature vectors and predicting interaction. The target property and drug property generate a features matrix for prediction.

IV. Databases Used in DTI Prediction

Interaction prediction demands the twin data items of drugs and targets, and a working knowledge of their interaction. The popular databases used in this study fall into two categories, drug-centered and target-centered. More than 20 databases associated with interaction prediction are not directly involved in DTI prediction, though the data contained therein maybe used as input for prediction. The popular database, KEGG, used here for prediction, is divided into the sub-databases of KEGG BRITE [46] and KEGG DRUG [47], incorporating a mass of biological data from genes and proteins.

A. Chemical European Molecular Biology Laboratory (ChEMBLdb)

The data gathered is a chemical database of bioactive molecules [48] which are collected from numerous literature studies. With millions of chemical compounds, 10,000 drugs and 12000 targets, the ChEMBLdb was established by the EMBL – European Bioinformatics Institute in 2002.

B. Chemical – Protein Annotation Resource (ChemProt)

ChemProt [49] has Chemical-Protein interactions data that integrates data from multiple databases of chemical protein annotations. It comprises data from the PDSP, DrugBank, PharmGKB, PubChem and STITCH databases. ChemProt also integrates therapeutic effects, adverse drug reactions and chemical-biological disease data.

C. Drug Gene Interaction Database (DGIdb)

This database has information on Druggable targets with their effects and drug-gene interaction data [50].

D. DrugBank

DrugBank is one of the most well-known databases in DTI study, with details about drug-like compounds, their different forms, target genes and side effects brought on by drug intake. The DTI data in this database that have been collected from an array of literature studies has extensive commercial uses [51].

E. Kyoto Encyclopedia of Genes and Genomes (KEGG)

KEGG is an out-of-the-box database with exhaustive details of genes and genome sequences [52]. The KEGG databases are divided into four categories. The first has three numbers of databases KEGG - BRITE, PATHWAY and MODULE. The second has four databases that carry genomic information– KEGG-GENOME, KEGG-GENE, KEGG-SSDB and KEGG ORTHOLOGY. The third has five databases with chemical information KEGG- COMPOUNDS, KEGG-REACTION, KEGG-RCLASS, KEGG-ENZYME and KEGG-GLYCAN. The fourth has four databases carrying health information– KEGG-DISEASE, KEGG-DRUG. The comprehensive KEGG has a wealth of DTI information and outclasses others.

F. Library of Integrated Network-Based Cellular Signatures (LINCS)

This database holds information on the KINOME scan. Kinases are small molecule-binding assays that help study the interaction between drug compounds for testing purposes. The database consists of 398 datasets on fluorescence imaging, ELISA and ATAC-sequence data [53].

G. PROMISCUOUS

The database has network-based drug repositioning data with information on drugs, proteins and the side effects of every drug. The information on protein is from the Unitprot database, while details on drugs and side effects are from the SuperDrug and Sider databases, incorporated into the *LINCS* [54].

H. Search Tool for Interacting Chemicals (STITCH)

STITCH has information on target or protein interaction with small molecules, collected from PubChem databases and literature studies [55].

I. SuperTarget

SuperTarget is a web resource that carries information on DTIs, drug metabolic rate, pathways, and Gene Ontology (GO) terms, as well as on adverse medical side effects. The DTI information is sourced from PubMed, DrugBank, KEGG, PDB and TTD, and potential drug-target relationships are extracted from Medline [56].

J. Therapeutic Target Database (TTD)

The Target Therapeutic Database has therapeutic information on protein and nucleic acid, assimilated from literature studies and miscellaneous databases with DTI data [57].

K. BRENDA -The Comprehensive Enzyme Information System (BRENDA)

This is an enzyme database with information on enzyme-ligand interaction. The data collected is drawn from literature studies based on enzyme nomenclature [58].

L. Drug Central

Drug Central is a Food and Drug Association (FDA)-approved drug database. The database incorporates relevant information on drugs in the form of structure, bioactivity and regulatory records, which are categorized as small molecule active ingredients and biological active ingredients [59].

M. Protein Drug Interaction Database (PDID)

Protein Drug Interaction Database (PDID) has DTI for all the structural proteome for human beings, with predictions made using the ILbind, SMAP and eFindSite software[60].

N. Pharos

Pharos is the user interface for giving knowledge about Illuminating Druggable Genome (IDG) to the knowledge management center for three of the protein families like GPCR, Ion Channel and Kinases [61].

O. PubChem

PubChem [62] has information about chemical substances and their biological activity. The PubChem database incorporates three databases–Substances, Compounds and BioAssay. The first stores data on chemical information, the second has exclusive chemical structures obtained from substances, while the third holds biological information on the extracted substances.

P. Super Drug

Super Drug [63] offers information on all drug features collected from several databases and incorporated here. The database has 2-Dimensional and 3-Dimensional structure information on small molecule drugs, side effects and drugs pharmacokinetics specifications.

Q. FDA Adverse Event Reporting System (FAERS)

The FDA Adverse Event Reporting System (FAERS) is a database with information obtained from adverse events and medication error reports submitted to the FDA on side effects, as well as keywords for drugs [64].

R. SIDe Effect Resource (SIDER)

SIDER is a database [65] that holds data on marketed medicines and their side effect information, including frequency of side effects, and also drug and its side effect classification.

S. International Union of Basic and Clinical Pharmacology (IUPHAR) / British Pharmacological Society (BPS) -The IUPHAR/BPS Guide to Pharmacology

The IUPHAR/BPS is considered as a guide to pharmacology [66] is an open access knowledge website that provides information on licensed drugs and their targets and holds information on small molecule drugs.

T. Cancer Drug Resistance Database (CancerDR)

CancerDR offers elaborate information on anti-cancer drugs and their pharmacological profiling. CancerDR helps in effective personalized cancer therapies and identifies gene-encoding drug targets, based on genetic and residual resistance [67].

U. Binding Database (DB)

Binding DB is a binding database that holds the DTI of small molecules as well as all the interaction data collected from an array of literature studies. This is an extensive database for protein ligand binding affinity [68].

V. ZINC is not Commercial (ZINC)

ZINC is the largest database [69] comprising every drug needed for new ligand discovery. Information on drugs and the targets they can interact with are collected here. ZINC is a major database for researchers looking for the chemical composition of their biological targets.

W. Psychoactive Drug Screening Program (PDSP)

The Psychoactive Drug Screening Program (PDSP) [70] screens compounds with previous reports of pharmacological, biochemical and behavioural activity. It is chiefly used to identify novel targets in the treatment of mental disorders.

X. A Summary of Databases

Table I, summarizes the general statistical information on every database.

TABLE I. Databases Involved in DTI Prediction

S. No	Databases	No. of Targets	No. of Drugs	No. of Interactions
1	ChEMBL	12482	1879206	15504603
2	ChemProt	20000	170000	-
3	DGI db	41100	9495	29783
4	DrugBank	5175	13338	26932
5	KEGG	19711	4948	260000
6	LINCS	1469	41847	-
7	PROMISCUOUS	6548	5258	23702
8	STITCH	9600000	430000	-
9	SuperTarget	6000	196000	330000
10	TTD	3101	34019	-
11	BRENDA	84000	20500	-
12	Drug Central	-	4543	-
13	PDID	3746	5100	-
14	Pharos	20244	130166	-
15	PubChem	79622	96157016	-
16	Super Drug	4456	4605	-
17	FAERS	-	24842	-
18	SIDER	1430	140064	-
19	IUPHAR/BPS	1396	1105	443
20	Cancer Dr	-	148	-
21	Binding DB	7020	489416	1132739
22	Zinc	-	20 million	-
23	PDSP	738	7449	-

TABLE II. Dataset Used in DTI Prediction

Dataset	Targets	Drugs	DTI
Enzyme	664	445	2926
Ion Channel	204	210	1476
GPCR	95	223	635
Nuclear Receptor	26	54	90

GPCR- G-Protein Coupled Receptor

V. Qualitative Analysis of Machine Learning Techniques for DTI Prediction

Qualitative analysis helps in an understanding of the ML techniques involved in DTI predictions, based on the quality and characteristics of the methods used. Qualitative analysis outcomes are descriptive, and inferences are drawn easily from the data obtained and the analysis of DTI prediction is shown in Table III-VII.

The Yaminishi et al. [71] Bench Mark (BM) dataset has been the only one used by many of the researchers for the purpose because it incorporates diverse drug and target data to create a new DTI dataset. The BM dataset is shown in Table II.

A. Review of Literature for Similarity-Based Methods

Similarity-based methods consider similarities between drugs and targets to identify DTIs. Perlman et al. [11] proposed a scheme that incorporates multiple drugs and targets similarity to predict DTI using the logistic regression SITAR (Similarity-based Inference of drug-TARgets) framework. Mei et al. [12] proposed a bipartite local model (BLM)-based method to handle the candidate problem of baseline BLM-NII (BLM with Neighbor-based Interaction profile Inferring). Van Laarhoven and Marichiori [13] developed a weighted nearest neighbor (WNN)algorithm that directly uses the GIP (Gaussian interaction profile) kernel by drawing up a profile of the interaction score for a new drug (WNN-GIP). Shi et al. [14] proposed a method to handle missing interactions using a cluster of similar targets that is Super

TABLE III. Qualitative Analysis of the Articles Using Similarity-Based Methods

Source	ML Tech	Dataset	Pre processing/ Feature Extraction	Feature Selection	Validation	Strength	Weakness	Outcome
Reference [11] (2011)	Logistic Regression (LR)	250 Proteins, 315 Drugs	-	Wrapper Feature Selection	10 Fold CV	Lists the selected features	Only 10 features are considered	Targets of 307 drugs are predicted
Reference [12] (2012)	Bipartite Local Model-NII (BLM-NII)	BM Dataset	-	-	LOOCV and 10 Fold CV	NII procedure for finding drugs and targets	Whenever new drug or target is given as input it is not considered as there is no training data	57 % of DTI has been predicted
Reference [13] (2013)	Weighted Nearest Neighbor (WNN)	BM Dataset	-	-	LOOCV and 5 Fold CV	Uses regularized least square algorithm to find the new drug based on the old drugs	No difference between indirect and direct targets. These are not measured to interact with drugs.	Prediction of DTI interaction which show top 5 prediction for each dataset.
Reference [14] (2015)	Super Target Clustering (STC)	BM Dataset	-	-	5 Fold CV	Finds missing interaction using cluster of targets.	Considers only about missing interaction not more about existing DTI	Finds new drugs and targets and potential interaction
Reference [15] (2016)	K-Nearest Neighbor (KNN)	BM Dataset	Finger print extraction for drugs	-	5 Fold CV, LOOCV	Hubness awareness and ensemble size gives high accuracy	LOOCV over fits and then shifted to 5 Fold CV	Improved prediction of DTIs around 12 prediction is found
Reference [16] (2017)	LPLNI	BM Dataset	-	-	LOOCV	Integrating similarities of different features	Considers only fingerprint as features for drugs	A promising tool for DTI prediction
Reference [17] (2017)	Multi-View DTI	1253drugs, 887 targets	-	-	20 trials of 5 Fold CV	Enrichment analyzes of drugs and targets	No details of experiments	56 newly identified clusters
Reference [18] (2018)	K-Nearest Neighbor (KNN)	BM Dataset	-	-	5 trials of 10 Fold CV	Calculating probability based weight and similarity based weight for targets	Considers only Ranking of top several integrations of drug and targets	34 % better prediction than previous methods

BM Dataset - Bench Mark dataset, CV- Cross Validation, LOOCV-Leave One Out Cross Validation.

TABLE IV. Qualitative Analysis of the Articles Using Matrix-Based Methods

Source	ML Tech	Dataset	Pre processing/ Feature Extraction	Feature Selection	Validation	Strength	Weakness	Outcome
Reference [19] (2009)	BRDTI	BM Dataset	-	-	5 Trials of 10 Fold CV	Incorporates target bias and context alignment for drug and target similarities	More survey based on DTI is to be done for better prediction.	DTI leads to Drug repurposing and adverse drug reaction prediction
Reference [20] (2012)	KBMF	BM Dataset	-	-	5 Fold CV	Interaction score is generated using factorization methods	Better for only 12 low dimensional projection	Similarity based DTIs.
Reference [21] (2017)	MLRE	608 protein, 326 drugs, 114 interactions	Structural view and chemical view of drug are extracted	-	5 Fold CV	Preserving the point wise linear regression	Noisy observation leads to disagreement data	Predict interaction based on chemical view with SVM and graph based methods
Reference [22] (2017)	VB-MK-LMF	BM Dataset	-	-	5 Trials of 10 Fold CV	DTI matrices are linked to weighted common observations	Works well for mid-sized datasets	DTI predicted
Reference [23] (2018)	Pseudo SMR	BM Dataset	Extraction of Pseudo AAC	-	5 Fold CV	Uses extremely randomized tree methods and it is computationally more efficient	Uses only Pseudo AAC Descriptors.	Predicted 15 Potential DTIs.

BM Dataset - Bench Mark Dataset, CV- Cross Validation, AAC –Amino Acid Composition.

TABLE V. Qualitative Analysis of the Articles Using Feature-Based Methods

Source	ML Tech	Dataset	Pre processing/ Feature Extraction	Feature Selection	Validation	Strength	Weakness	Outcome
Reference [25] (2011)	Regularized Least Square	BM Dataset	-	-	LOO CV and 5 Trials of 10 Fold CV	Combining GIP with target kernel and drug kernel	Increase kernel with more information about DTI	15 known interaction was predicted
Reference [26] (2016)	Krons-Regularized Least Square	BM Dataset	Replaced missing values with mean of data	-	5 Fold CV	Incorporates both known and unknown interaction and make a general purpose learner	Balancing the data is not considered	Prediction of interval as measure of confidence
Reference [27] (2016)	Weighted SVM	BM Dataset	Structural similarity, Gene Function similarity was extracted	-	5 Fold CV	Finds some unlabeled sample as negative sample and also considers positive samples beneath unlabeled samples	Asks for using structure but we cannot get structure for all the targets	Predicts Interaction and listed 3 top known interaction
Reference [28] (2016)	Ensemble learning	5877Drugs 3348Targets 12674DTI	PROFEAT for Target and Rcpi for Drug	-	5 Fold CV	Ensemble learning to address issues of class imbalance	Oversampling is done which increases noise	Predicted more than 20 Known DTI
Reference [29] (2017)	Discriminate Vector Machine	BM Dataset	AAC feature were Extracted	Principal Component Analysis (PCA)	5 Fold CV	Uses LBP histogram vectors which retains evolutionary information of amino acid	Only AAC information is used for prediction	Not listed the predicted DTI
Reference [30] (2017)	Support Vector Machine	BM Dataset	-	-	10 Fold CV	Multiple Kernel combination is used for prediction	GIP based prediction	Compound-Protein-Interaction
Reference [31] (2017)	REP Tree Algorithm	2719 E 1372 IC 630 GPCR 86 NR	-	-	10 Fold CV	Considers different families of proteins by using various learning rate	No cross validation is done	DTI prediction
Reference [32] (2017)	Adaboost	BM Dataset	PSSM for target and SMILE for drug were extracted	Sequential Forward Feature Selection (SFFS)	5 Fold CV	Balanced Data using RUS and CUS techniques	Not considered domain features	Listed top 10 known interaction
Reference [33] (2018)	Bagging based ensemble	5877Drugs 3348Targets 12674 DTI	PROFEAT for Target and Rcpi for Drug	-	10 Fold CV	Considered class imbalance and used Neighbourhood balanced bagging for balancing the data and active learning strategy is used	Not discussed about Features	14 out of 16 known interactions have been detected.

BM dataset - Bench Mark dataset, CV- Cross Validation, LOOCV-Leave One Out Cross Validation, PROFEAT-PROtein FEATures, AAC- Amino Acid Composition, Rcpi-R package for extracting features for compound protein interaction..

Target Clustering (STC). Buza K [15] proposed a K-nearest neighbor (KNN)-based method with hubness-aware classification and error correction to maximize the detrimental effect of bad hubs (EcKNN-KNN with error correction). Zhang et al. [16] posited a framework that develops a drug-drug linear neighbourhood, calculates the similarities, and predicts drug-target interaction profile and label propagation (LPLNI-Label Propagation with Linear Neighbourhood Information). Zhang et al. [17] developed a clustering algorithm by incorporating drug and target data from structural and chemical viewpoints with existing knowledge of interactions (MDTI- Multiview DTI). Shi and Li [18] advanced an improved Bayesian ranking DTI method that adds weights for unknown drugs and targets using weighted neighboring drugs and targets (WBRDTI–Weighted Bayesian Ranking DTI).

B. Review of Literature for Matrix-Based Methods

Matrix-based methods use matrix similarity for DTI prediction. Rendle et al. [19] proposed an algorithm based on the Bayesian Personalized Ranking (BPR) matrix factorization which incorporates drug and target similarities to predict DTIs (BPRDTI). Gonen [20] proposed a method to factorize the matrices with interaction score matrix so as to find new drugs and targets and determine their interaction using kernelized Bayesian matrix factorization (KBMF). Li et al. [21] introduced an algorithm to find a low-rank representation embedding (LRE) technique and fix errors in point wise linear reconstruction. This was done to obtain a different view of the structural and chemical features of drugs and targets as Single view

TABLE VI. Qualitative Analysis of the Articles Using Network-Based Methods

Source	ML Tech	Dataset	Pre processing/ Feature Extraction	Feature Selection	Validation	Strength	Weakness	Outcome
Reference [34] (2012)	Network based Inference (NBI)	BM Dataset	-	-	10 Fold CV	Used a bipartite graph for prediction	Imbalanced data is used	5 new DTI were predicted
Reference [35] (2012)	Network-based Random Walk with Restart on the Heterogeneous network (NRWRH)	BM Dataset	-	-	LOOCV	Used RWR to get potential DTI using bipartite graph network	Leaves the target which has no drug it is considered ass zero matrix	29 new DTI were predicted
Reference [36] (2013)	Network-Consistency-based Prediction Method (Net CBP)	BM Dataset	-	-	Not discussed Properly	DTI predicted using bipartite graph network	Considered as zero matrix	Listed out several DTI
Reference [37] (2015)	Normalized Multi information Fusion	BM Dataset	-	-	Not discussed properly	Integrates robust PCA with biological information	In order to improve performance more negative dataset to be built to find the interactions.	Predicts interaction
Reference [38] (2015)	Random Walk Restart (RWR)	467Targets 544Drugs	-	-	-	RWR on heterogeneous network using chemical features	Considered only fingerprints features for drugs	110 drugs predicted for 3419 targets
Reference [39] (2018)	IN - Random Walk with Restart (RWR)	12015 Drug 1895445 Target	-	Principal Component Analysis (PCA)	5 Fold CV	Used both labelled and unlabeled data for prediction	Data is imbalanced	Predicts interaction between drug and targets
Reference [40] (2019)	Neighbourhood Regularized Logistic Matrix Factorization (NRLMF)	BM Dataset	Calculates similarities of drugs and targets	-	10 Fold CV	Improved using rescoring matrix	Not more parameters are considered	Predicts interaction but not listed

BM Dataset - Bench Mark Dataset, CV- Cross Validation, LOOCV-Leave One Out Cross Validation.

LRE and Multiview LRE, respectively (LRE). Bolgar et al. [22] developed a method integrating multiple kernels, weights, and graphs, all regularized to model the probability of DTI prediction (VB-MK-LMF). Huang et al. [23] propounded an extension of the structure activity relationship classification by implementing the extremely randomized tree (ERT) using the pseudo substitution matrix representation (SMR) of the target (Pseudo-SMR). Marta et al. [24] proposes a local model-agnostic for interaction prediction.

C. Review of Literature for Feature-Based Methods

Feature-based methods consider drug and target features for DTI prediction. Van Laarhoven et al. [25] proposed an algorithm that integrates the DTI network information with the Gaussian Interaction Profile kernel using the Regularized Least Square (RLS). Ezzat et al. [26] developed a framework for DTI prediction using the voting of the decision tree, random forest, STACK and Laplacian Eigen base classifiers, and also considered imbalanced classes for prediction. Nascimento et al. [27] advanced a method that incorporates both known and unknown interaction data using the RLS. Lan et al [28] developed a framework for DTI prediction by taking unlabeled samples using the weighted SVM (PUDT-Positively Unlabeled Drug Targets). Li et al. [29] proposed a method to find DTIs as a structure activity relationship (SAR) classification with the principal component analysis (PCA), using the Discriminative Vector Machine (DVM). Ohue et al. [30] proposed an approach that uses virtual screening and the Pairwise Kernel Method (PKM). Zhang et al. [31] proposed an ensemble-based approach for a random projection ensemble (RPE) of the REP tree algorithm (Drug RPE). Rayhan et al. [32] developed a model using targets in the form of a matrix (position-specific scoring matrix - PSSM) and drug molecules features for DTI prediction using the AdaBoost classifier (iDTI-EsBoost). Sharma and Rani [33] proposed an ensemble (Bagging-Ensemble) model that uses active learning methodology to predict DTIs (BE-DTI).

D. Review of Literature for Network-Based Methods

These methods use networks of similar drugs and targets for DTI prediction. Cheng et al. [34] proposed a bipartite Network Based Inference (NBI) method for DTI prediction. Chen et al. [35] developed an RWR framework to get potential DTIs using a bipartite graph network (NRWRH-Network-based Random Walk with Restart on Heterogenous network). Chen et al. [36] used this method for both labelled and unlabeled data DTI prediction (NETCBP-Network Consistency-based Prediction). Peng et al. [37] proposed a method that incorporates the PCA to reduce dimensions and integrate data from multiple drug and target sources for DTI prediction (NMIF-Normalized Multi-Information Fusion). Seal et al. [38] proposed a model that needs matrix inversion and score of relevance between two nodes in a weighted graph of DTIs (RWR-Random Walk with Restart). Huang et al. [39] proposed a 2-network-based rank algorithm that involves the random walk and bipartite graph (IN-RWR-intra network with Random Walk). Ban et al. [40] developed a method based on improving the NRLMF algorithm by calculating the NRLMF scores as the expected beta distribution values.

TABLE VII. Qualitative Analysis of the Articles Using Deep Learning-Based Methods

Source	ML Tech	Dataset	Pre processing/ Feature Extraction	Feature Selection	Validation	Strength	Weakness	Outcome
Reference [41] (2017)	Deep DTI	1520 Targets 1412 Drugs 12524 samples	-	-	10 Fold CV	Uses DBN and Fine tune RBM in greedy way.	Only known interaction are used	DTI probability which are useful for drug repurposing
Reference [42] (2018)	Deep DTA	442 Targets 68 Drugs 30056 DTI	-	-	Concordance index	Creating CNN blocks of targets, drugs	Predefined features are considered for CNN blocks of protein	Predicts binding affinity
Reference [43] (2018)	AUTO DNP	BM Dataset	PSSM for Target and PubChem fingerprint has taken for drugs	-	5 Fold CV	Uses Auto encoder blocks to create Deep NN	Only CTD descriptors are considered.	Predicts interaction
Reference [44] (2019)	LASSO – DNN	3546 Proteins 5834 Drugs 14792 DTI	-	-	10 Fold CV	Considers Tripeptide composition feature of proteins	More number of functions are used.	Diseases treated by drug and its association with breast cancer is listed
Reference [45] (2019)	Deep Convolution-DTI	3675Targets 11950Drugs 32,568 DTI	-	t-distributed stochastic neighbor embedding (t-SNE)	5 Fold CV	Similarity acts as a informative descriptors	Considers only CTD descriptors of targets	Predicts interaction

BM Dataset - Bench Mark Dataset, CV- Cross Validation, CTD – Composition, Transition and Distribution, PSSM - Position Specific Scoring Matrix, PubChem – PubChem is a Chemical Information database.

Beta distribution value is calculated using the interaction information and NRLMF score (NRLMF-beta).

E. Review of Literature for Deep Learning-Based Methods

Deep learning-based methods use the drug and target features for DTI prediction. Wen et al. [41] proposed a method that takes raw target and drug features using a deep belief network (DBN) and predicts DTI in drugs approved by the Food and Drug Association (DeepDTIs). Ozturk et al. [42] proposed a DTI prediction model using target sequences and drug molecule to predict drug target binding affinity (DeepDTA). Wang et al [43] developed a computational model using a stacked auto encoder for DTI prediction (AUTO-DNP). You et al. [44] presented a method based on protein and drug features with LASSO regression model in tandem with the deep neural network (DNN) to predict DTI (LASSO-DNN). Lee et al. [45] proposed a DTI prediction model using local protein residue patterns in DTI (DeepConv-DTI).

VI. Quantitative Analysis of Machine Learning Techniques in DTI Prediction

Quantitative analysis is applied to determine the best prediction performance method, using different ML techniques with appropriate metrics. The prediction method must deal with the steps of data pre-processing and feature selection, as well as drug and target integration. The best machine learning prediction method includes the hyper parameters and association index for DTI prediction. Of the various ML techniques [11]-[44] available, the best is chosen for prediction. Tables X-XIV depict the quantitative analysis of the results of several ML methods in DTI prediction that help enhance performance.

A. Performance Metrics

A confusion matrix is used to calculate performance measures from test set values in terms of true positives, true negatives, false positives and false negatives among classes that are to be classified as integrates or not integrates. Table VIII shows the confusion matrix for DTI and Table IX the performance metrics used. Integrates here refers to drugs that produce a positive DTP result, that is, the integrating drug can be used to treat a target it integrates with. The converse is true with non integrates, which refers to drugs that produce a negative DTP result, that is, the non integrating drug cannot be used to treat a target it does not integrate with.

TABLE VIII. Confusion Matrix

	Integrates	Non Integrates
Integrates	True Positive	False Positive
Non integrates	False Negative	True Negative

TABLE IX. Performance Metrics Used in DTI Prediction

S. No	Metrics Used	Formula	Metrics Description
1.	Accuracy	$(TP+ TN)/(TP+TN+FP+FN)$	Accuracy is the ratio of correct prediction out of total number of predictions
2.	Sensitivity/ Recall	$TP/(TP+FN)$	Measure of quantity
3.	Precision	$TP/(TF+FP)$	Measure of quality
4.	AUC	False Positive vs. True Positive	Curve shows the relation between False Positive and True Positive
5.	AUPR	Precision vs. Recall	Curve shows the relationship between the Precision and Recall
6.	MCC	$\frac{(TP * TN) - (FP * FN)}{\sqrt{((TP+FP)(TP+FN)(TN+FP)(TN+FN))}}$	Mathew's Correlation Coefficient
7.	F1 Score	$TP/(TP+1/2+TP/(FP+FN))$	Harmonic average of Precision and Recall

TABLE X. Quantitative Analysis of the Similarity-Based Methods Used in DTI Prediction

S. No	ML Tech.	Similarity Based Methods																			
		Accuracy				Sensitivity/ Recall				Precision/nDCG				AUC				AUPR/MAP			
		E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N
1.	LR	-	-	-	-	-	-	-	-	-	-	-	-	92.2	92.7	94.6	86.3	87.7	88.9	93.9	85.1
2.	BLM-NII	-	-	-	-	-	-	-	-	-	-	-	-	98.8	99.0	98.4	98.1	92.9	95.0	86.5	86.6
3.	WNN	-	-	-	-	-	-	-	-	-	-	-	-	81.9	75.5	84.8	78.8	29.9	24.9	30.8	43.4
4.	STC	-	-	-	-	-	-	-	-	-	-	-	-	81.2	81.1	87.5	87.1	38.5	36.7	41.4	53.3
5.	KNN	-	-	-	-	-	-	-	-	-	-	-	-	95.4	97.2	97.2	-	83.7	85.5	62.8	-
6.	LPLNI	-	-	-	-	-	-	-	-	-	-	-	-	97.0	97.6	99.4	99.1	90.6	94.6	96.8	94.9
7.	Multi-view DTI	-	-	-	-	-	-	-	-	-	-	-	-	86.9				-	-	-	-
8.	KNN	-	-	-	-	-	-	-	-	nDCG 90.8	95.9	94.0	94.5	98.3	98.4	96.2	94.8	MAP 88.0	94.2	91.5	92.7

E-Enzyme, IC-Ion Channel, G-G-Protein Coupled Receptor (GPCR), N-Nuclear Receptor, AUC-Area Under Curve, AUPR-Area Under Precision Recall, nDCG-normalized Discounted Cumulative Gain PPV-Positive Predicted Values, MCC-Mathew's Correlation Coefficient, MAP-Mean Average Precision.

TABLE X. Quantitative Analysis of the Matrix-Based Methods Used in DTI Prediction

S. No	ML Tech.	Matrix based Methods																			
		Accuracy				Sensitivity/Recall				Precision/nDCG				AUC				AUPR/MCC			
		E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N
1.	BRDTI	-	-	-	-	-	-	-	-	nDCG 89.7	95.3	92.9	94.8	98.1	98.2	95.5	92.3	-	-	-	-
2.	KBMF	-	-	-	-	-	-	-	-	-	-	-	-	83.2	79.9	85.7	82.4	-	-	-	-
3.	MVLRE	-	-	-	-	-	-	-	-	-	-	-	-	65.0	51.4	61.7	-	-	-	-	-
4.	VB-MK LMF	-	-	-	-	-	-	-	-	-	-	-	-	98.7	98.9	97.6	95.7	89.0	91.0	80.0	77.0
5.	Pseudo SMR	89.4	87.8	82.9	83.3	89.5	87.9	82.1	95.2	90.2	87.8	82.1	76.3	96.0	93.8	90.5	96.3	MCC 81.8	78.7	71.8	71.6

E-Enzyme, IC- Ion Channel, G- G-Protein Coupled Receptor (GPCR), N- Nuclear Receptor, AUC-Area Under Curve, AUPR- Area Under Precision Recall, PPV-Positive Predicted Values, MCC- Mathew's Correlation Coefficient, nDCG-normalized Discounted Cumulative Gain.

TABLE XII. Quantitative Analysis of the Feature-Based Methods Used in DTI Prediction

S. No	ML Tech.	Feature Based Methods																			
		Accuracy/PPV/MCC/ F1 Score				Sensitivity/Recall				Precision				AUC				AUPR/MCC/ F1 Score			
		E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N
1.	RLS	-	-	-	-	-	-	-	-	-	-	-	-	98.2	98.5	94.5	88.7	88.1	91.8	70.0	60.4
2.	Krons-RLS	-	-	-	-	-	-	-	-	-	-	-	-	97.9	98.7	95.1	92.4	-	-	-	-
3.	Weighted SVM	PPV 36.0	74.0	58.0	64.0	24.0	14.0	16.0	7.0	99.0	99.0	94.0	97.0	88.4	83.1	87.8	88.5	-	-	-	-
4.	Ensemble Learning	-	-	-	-	-	-	-	-	-	-	-	-	90.0				-	-	-	-
5.	DVM	93.1	91.7	89.3	92.2	92.9	92.6	89.2	96.6	93.1	90.9	89.4	88.6	92.8	91.7	88.56	93.00	MCC 86.3	83.4	78.77	84.80
6.	REP Tree	94.0	91.0	88.0	88.0	92.0	89.0	81.0	87.0	90.0	86.0	83.0	79.0	98.0	97.0	94.0	93.0	F1 Score 91.0	88.0	82.0	83.0
7.	Adaboost	MCC 18.0 F1 Score 10.0	29.0 20.0	26.0 19.0	22.0 24.0	85.0	84.0	84.0	87.0	85.0	78.0	80.0	92.0	96.0	93.0	93.0	92.0	68.0	48.0	50.0	79.0
8.	BE-DTI	-				88.0				-				92.7				88.6			

E-Enzyme, IC- Ion Channel, G- G-Protein Coupled Receptor (GPCR), N- Nuclear Receptor, AUC-Area Under Curve, AUPR- Area Under Precision Recall, PPV-Positive Predicted Values, MCC- Mathew's Correlation Coefficient.

TABLE XIII. Quantitative Analysis of the Network-Based Methods Used in DTI Prediction

Network Based Methods

S. No	ML Tech.	Accuracy				Sensitivity/Recall				Precision				AUC				AUPR			
		E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N
1.	NBI	-	-	-	-	93.5	98.1	94.8	85.1	97.5	97.6	94.6	83.8	-	-	-	-	-	-	-	-
2.	NRWRH	-	-	-	-	85.0	-	-	-	99.0	-	-	-	-	-	-	-	-	-	-	-
3.	Net CBP	-	-	-	-	-	-	-	-	-	-	-	-	82.5	80.3	82.3	83.9	-	-	-	-
4.	NMIF	-	-	-	-	-	-	-	-	-	-	-	-	83.0	82.0	82.0	80.0	81.0	78.0	74.0	71.0
5.	RWR	-	-	-	-	-	-	-	-	-	-	-	-	70.9				-	-	-	-
6.	IN-RWR/ Corank	82.2				-	-	-	-	-	-	-	-	95.1				-	-	-	-
7.	NRLMF-beta	-	-	-	-	-	-	-	-	-	-	-	-	99.0	99.0	97.5	96.4	89.7	91.3	75.5	75.5

E-Enzyme, IC- Ion Channel, G- G-Protein Coupled Receptor (GPCR), N- Nuclear Receptor, AUC-Area Under Curve, AUPR- Area Under Precision Recall, PPV-Positive Predicted Values, MCC- Mathew's Correlation Coefficient.

TABLE XIV. Quantitative Analysis of the Deep Learning-Based Methods Used in DTI Prediction

Deep Learning based Methods

S. No	ML Tech.	Accuracy				Sensitivity/Recall				Precision				AUC				AUPR/MCC			
		E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N
1.	Deep DTI	85.8				82.2				-				91.5				-			
2.	Deep DTA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	71.4			
3.	AUTO DNP	94.1	91.1	86.6	80.5	95.5	95.6	81.6	76.2	92.9	87.7	91.0	84.1	94.2	91.0	87.4	81.7	MCC 88.3	82.7	73.9	61.8
4.	LASSO-DNN	81.0				-	-	-	-	-	-	-	-	89.0				-	-	-	-
5.	Deep Convolution DTI	75.0				85.0				70.0				80.0				-	-	-	-

E-Enzyme, IC- Ion Channel, G- G-Protein Coupled Receptor (GPCR), N- Nuclear Receptor, AUC-Area Under Curve, AUPR- Area Under Precision Recall, PPV-Positive Predicted Values, MCC- Mathew's Correlation Coefficient.

VII. Discussion

The analysis shows that the chemogenomics-based approach to DTI prediction is ideally suited to interaction prediction. A review of the qualitative and quantitative analyses offers an overview of the dataset, preprocessing, feature selection techniques, validation and ML classification techniques used in DTI prediction, all of which are discussed in this section.

A. The Dataset

The benchmark Yaminishi et al. dataset [71] is invariably used in DTI prediction, with its four enzyme (E), ion channel (IC), G-protein coupled receptor (GPCR) and nuclear receptor (NR) classes and the DTI positive pairs of each class. Apart from the benchmark dataset above, others are used as well [11], [17], [21], [26], [31]. Deep learning-based prediction works with more dynamic data. An attempt has been made in [44] to construct a negative DTI dataset, which is significant in that it facilitates the assimilation of targets not taken into the prediction process. The number of instances used, which ranges from 250 to 5500, may be increased or decreased, depending on the purpose of the research.

B. Preprocessing and Balancing Techniques

Major issues in DTI prediction are brought on by the data obtained from miscellaneous sources, which may have a different range of values or none at all. Missing values from known data are inferred, based on the observed values in the data structure. Preprocessing techniques are, generally speaking, not used on the data because they are curated when collected from different sources. When the data are incorporated, however, values may go missing or are replaced, and there is thus a need for preprocessing. The preprocessing employed in [26] to replace missing values uses the mean values of the data. Employing preprocessing techniques like data cleaning enhances the quality of the data for further processing.

From the qualitative analysis tables III-VII, it is found that the dataset used in the prediction process is unbalanced and may affect the performance of the classifiers. Balancing techniques include balancing the data using oversampling [26], [32], [33], though it increases negative outcomes. For DTI prediction, undersampling can be suggested to improve the positive outcomes.

C. Feature Extraction Methods

Feature Extraction is done to reduce the dimensionality of the input features by creating a new set of features from the original features which gains the important features of the data and also reduces the dimension of the features, which increases the speed of learning and generalization of machine learning. It can also be done through various tools available for it. In drug discovery researchers use several tools for feature extraction, the trending tools are PROFEAT and Protr for protein feature extraction, Rcpi and PADEL Descriptor for drug feature extraction. The research work which uses these tools for feature extraction are [28], [33].

D. Feature Selection Methods

Feature selection is of fundamental importance, because the extracted features increase data dimensions and result in problems with over fitting. Feature selection techniques reduce the number of features by selecting the most important ones from the given input. It is clear from the analysis that target features can be categorized into three –structural, evolutionary and sequence. While the drug feature is structural, the number of target features considered varies from 1080 to 1498. Likewise, drug features vary, depending on whether they are 1D or 2D and on the fingerprint of the drugs selected. Tables III-VII in [11]-[18] that showcase similarity-based methods only consider similarities between drug-drug, target-target and drug-target for DTI prediction, which means that only similar drugs interact with similar targets. So in similarity based methods, drug-based and target-based features are considered unimportant for DTI prediction. Further, similarity-based methods do not handle large-scale datasets. Matrix-based methods [19]-[23] consider only drug and target similarities, and no other features are taken for prediction. Also, matrix-based methods only handle small-scale datasets. Of the feature-based methods used in [25]-[33], the Sequential Forward Feature Selection (SFFS) technique is applied in [33], where the different feature sets considered are added sequentially, one by one, to evaluate the dataset. It is observed that the structural feature, which is one of the most influential target features, plays a significant role in DTI prediction, and may vary with the dataset taken. Finding the most influential features is important to feature selection. The network-based methods in [34]-[40] take different sets of features and handle them appropriately by selecting the most important drug and target features. Compact feature learning is undertaken in [39] by applying the Diffusion Component Analysis (DCA), which constructs a low-dimensional vector representation for each drug and target using diffusion distribution. It helps find the best interpretable features. The deep learning-based methods discussed [41]-[45] use the t-distributed Stochastic Neighbor Embedding (t-SNE) technique to reduce input feature dimensionality. Deep learning-based methods consider dynamic data and dynamic features. The Convolution Neural Network (CNN) used in [45] handles features with ease and finds the most potent ones. Given that deep learning-based methods deal with large-scale datasets well, future research that applies deep learning will execute DTI prediction better.

E. Validation Methods

The qualitative analysis depicts that the 10-fold Cross-Validation (CV) and 5-fold cross-validation offer better results than other CV techniques like the Leave-One-Out CV (LOOCV) and jackknife. Approaches using the LOOCV have problems with over fitting. DTI predictions are evaluated using AUC and AUPR values. The AUC values of the classifiers show better results when the 10-fold CV is used to validate the methods. AUC is chosen because it distinguishes between classes and validates the model's capacity even when the dataset is imbalanced.

F. ML Techniques Engaged in DTI Prediction

The qualitative analysis table III-VII, depicts the various classifiers used, one outclasses the rest at DTI prediction. Ranking algorithms like Bayesian ranking are used to rank DTI [20]. The SVM [19], [22] classifier, which handles target and drug features by calculating them separately and reducing prediction complexity cannot determine the relationship between the features and may produce a large number of false positives. The KNN [18], [20] falls short, performance-wise, in its inability to handle features and large-scale datasets. Ensemble learning [27] handles large-scale and high-dimensional data. The Adaboost classifier separates the data and classifies them to get the most appropriate features [32]. The decision tree manages missing data thoroughly and uses diversity to learn features based on instances for improved accuracy [33]. Logistic regression [11], [16] operates data integration strategies effectively. The DVM [29] influences features strongly in its handling of outliers. As far as feature-based methods are concerned, the random forest outperforms the rest, while the regularized least square (RLS) performs well in tandem with more influential features. In terms of performance, the WBR-DTI, VB-MK-LMF, NRLMF-beta and CNN find the best features for DTI prediction.

From the quantitative analysis table X-XIV, the progress made is evaluated using AUC values, with marked improvements in the SVM from 61.7% [19] to 96.34% [22], the KNN from 92.3% [18] to 95.4% [20], and LR from 85.1% [11] to 95.32% [16]. Among the classifiers used in DTI prediction, the SVM gives the best prediction results with an improvement of 34.64%. The random forest and decision tree used in ensemble learning give an AUC value of 90%. Adaptive Boosting and RLS give AUC values of 88.7% and 97%, respectively. The WBR-DTI and VB-MK-LMF give an AUC value of 98%, while the NRLMF-beta gives 96%.

However, the results are based on the data given as input. The new model developed may perform poorly, with imbalanced data and missing values. The qualitative analysis tables III-VII show that the dataset has more negative than positive predictions, owing to the nature of the dataset used for DTI prediction. The quantitative analysis tables X-XIV depict that matrix factorization-based methods perform best for DTI prediction, though deep learning-based methods handle large-scale data and find the most influential features and some of the papers gives light to other process like detecting adverse reaction of drugs [72]. This review has thus laid out a thorough understanding of datasets, feature selection methods and validations, as well as a comparison of the classifiers used for DTI prediction

VIII. Conclusion and Future Scope

It is concluded from the review that much research has focused chiefly on chemogenomics, and this is because DTI based on drug and target features and similarities may be found without their structures. The method works well by finding the most influential features using a range of classifiers for DTI prediction. The classifiers use only known static interaction for training the model, given that the interaction data is static. Though static data has largely been used as a benchmark dataset for interaction prediction, dynamic data may be considered so the problem of new DTI is resolved. Several studies have only considered target features (like the AAC, CTD and pseudo AAC) and the PubChem fingerprint for drugs. There are, therefore, plenty of research opportunities to predict drugs using the influence of all the features. Influential features may vary from one technique to another. There is, however, a delay in finding influential features, since one feature may not be as important for prediction as another. More data are to be considered for finding the most influential features, which is possible with the introduction of big data for prediction. The ML techniques used by the deep learning-based and matrix-based methods were found to predict DTI better than others. It is recommended, considering the above, that future researchers focus on building a negative dataset for interaction prediction. Feature scaling or feature engineering techniques may be applied to enhance the dataset. New databases can be created by collecting data from numerous sources and incorporating appropriate parameters or influential features for future research. Further, future models developed for DTI prediction must consider every feature for drug prediction. The model developed, based on ML techniques, should be able to update information on drugs and targets constantly for new interaction prediction. Thus, the model must be able to predict interaction, based on prior knowledge, without having to be trained on every occasion. Such a model is likely to offer the best interaction prediction.

References

[1] L. Martin, M. Hutchens, C. Hawkins, A. Radnov, “How much do clinical trials cost?,” Nature Reviews-Drug Discovery, vol. 16, no. 6, pp. 381-382, June 2017, DOI: 10.1038/nrd.2017.70.

[2] S.J. Swamidass, “Mining small-molecule screens to repurpose drugs,” Briefing in Bioinformatics, vol. 12, no. 4, pp. 327–335, 2011, DOI: 10.1093/bib/bbr028.

[3] F. Moriaud, S.B. Richard, S.A. Adcock, L. Chanas-Martin, J.S. Surgand, M. Ben Jelloul, F. Delfaud, “Identify drug repurposing candidates by mining the protein data bank,” Briefings in Bioinformatics, vol. 12, no. 4, pp. 336-340, Jul 2011, DOI: 10.1093/bib/bbr017.

[4] R. Chen, X. Liu, S. Jin, J. Lin and J. Liu, “Machine learning for drug-target interaction prediction,” Molecules, vol. 23, no. 9, pp. 2208, 2018, DOI: 10.3390/molecules23092208.

[5] T.T. Talele, S.A. Khedkar and A.C. Rigby, “Successful Applications of Computer Aided Drug Discovery: Moving Drugs from Concept to the Clinic,” Current Topics in Medicinal Chemistry, vol. 10, no. 10, pp. 127, 2010, DOI: 10.2174/156802610790232251.

[6] T. Usha, D. Shanmugarajan, A.K. Goyal, C.S. Kumar and S.K. Middha, “Recent updates on computer-aided drug discovery: time for a paradigm shift,” Current topics in medicinal chemistry, vol. 17, no. 30, pp. 3296-3307, 2017, DOI: 10.2174/1568026618666180101163651.

[7] L. Jacob, J-P. Vert, “Protein-ligand interaction prediction: an improved chemogenomics approach,” Bioinformatics, vol. 24, no. 19, pp. 2149–2156, 2008, DOI: 10.1093/bioinformatics/btn409.

[8] D. Rognan, “Chemogenomic approaches to rational drug design”, British Journal of Pharmacology, vol. 152, no. 1, pp. 38–52, 2007, DOI:10.1038/sj.bjp.0707307.

[9] A. Nath, P. Kumari, R. Chaube, “Prediction of human drug targets and their interactions using machine learning methods: current and future perspectives,” Methods in molecular biology, Springer, NY, USA, vol. 1762, pp. 21–30, 2018, DOI:10.1007/9781493977567_2.

[10] L. Lü, T. Zhou, “Link prediction in complex networks: a survey”, Physica A, vol. 390, pp. 1150–1170, 2011, DOI: 10.1016/j.physa.2010.11.027.

[11] L. Perlman, A. Gottlieb, N. Atias, E. Ruppin, R. Sharan, “Combining drug and gene similarity measures for drug-target elucidation,” Journal of computational biology: a journal of computational molecular cell biology, vol. 18, no. 2, pp. 133–145, 2011, DOI:10.1089/cmb.2010.0213.

[12] J.-P. Mei, C.-K. Kwoh, P. Yang, X.L. Li, J. Zheng, “Drug–target interaction prediction by learning from local information and neighbours”, Bioinformatics, vol. 29, no. 2, pp. 238–245, 2012, DOI: 10.1093/bioinformatics/bts670.

[13] T. Van Laarhoven, E. Marchiori, “Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile”, PloS One, vol. 8, no. 6, pp. e66952, 2013, DOI: 10.1371/journal.pone.0066952.

[14] J.-Y. Shi, S.-M. Yiu, Y. Li, H.C. Leung, F.Y. Chin, “Predicting drug–target interaction for new drugs using enhanced similarity measures and super-target clustering,” Methods, vol. 83, pp. 98–104, 2015, DOI: 10.1016/j.ymeth.2015.04.036.

[15] K. Buza, “Drug–target interaction prediction with hubness aware machine learning,” In: 2016 IEEE 11th International Symposium on Applied Computational Intelligence and Informatics (SACI), IEEE, New York, USA, 2016, pp. 37–40, DOI: 10.1109/SACI.2016.7507416.

[16] W. Zhang, Y. Chen, D. Li, “Drug–target interaction prediction through label propagation with linear neighborhood information,” Molecules, vol. 22, no. 12, pp. 2056, 2017, DOI: 10.3390/molecules22122056.

[17] X. Zhang, L. Li, M.K. Ng, S. Zhang, “Drug–target interaction prediction by integrating multiview network data”, Computational Biology and Chemistry, vol. 69, pp. 185–193, 2017, DOI: 10.1016/j.compbiolchem.2017.03.011.

[18] Z. Shi, J. Li, “Drug–target interaction prediction with weighted Bayesian ranking,” In: Proceedings of the 2nd International Conference on Biomedical Engineering and Bioinformatics, ACM, London, United Kingdom, 2018, pp. 19–24.

[19] S. Rendle, C. Freudenthaler, Z. Gantner, Z. Gartner, L. Schmidt-Thieme, “BPR: Bayesian Personalized Ranking from implicit feedback,” In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press, McGill, Canada, 2009, pp. 452–461, DOI: 10.1145/3278198.3278210.

[20] M. Gönen, “Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization,” Bioinformatics, vol. 28, no. 18, pp. 2304–2310, 2012, DOI:10.1093/bioinformatics/bts360.

[21] L. Li, M. Cai, “Drug target prediction by multi-view low rank embedding”, IEEE/ACM Transactions on Computational Biology and Bioinformatics vol. 16, no.5, pp.1712-1721, 1 Sep-Oct 2019, DOI: 10.1109/TCBB.2017.2706267.

[22] B. Bolgár, P. Antal, “VB-MK-LMF: fusion of drugs, targets and interactions using variational Bayesian multiple kernel logistic matrix factorization”, BMC Bioinformatics, vol. 18, no. 1, pp. 440, 2017, DOI: 10.1186/s12859-017-1845-z.

[23] Y.A. Huang, Z.H. You, X. Chen, “A Systematic Prediction of Drug-Target Interactions Using Molecular Fingerprints and Protein Sequences”, Current protein & peptide science, vol. 19, no. 5, pp. 468-478, 2018, DOI: 10.2174/1389203718666161122103057.

[24] M. Caro-Martínez, G. Jiménez-Díaz, J. A. Recio-García. “Local Model-Agnostic Explanations for Black-box Recommender Systems Using Interaction Graphs and Link Prediction Technique”, International Journal of Interactive Multimedia and Artificial Intelligence, 2021, DOI: 10.9781/ijimai.2021.12.001.

[25] T. Van Laarhoven, S.B. Nabuurs, E. Marchiori, “Gaussian interaction profile kernels for predicting drug–target interaction”, Bioinformatics, vol. 27, no. 21, pp. 3036–3043, 2011, DOI: 10.1093/bioinformatics/btr500.

[26] A. Ezzat, M. Wu, X.-L. Li, C.-K. Kwoh, “ Drug–target interaction prediction via class imbalance-aware ensemble learning”, BMC Bioinformatics, vol. 17, no. 19, pp. 509, 2016, DOI:10.1186/s12859-016-1377-y.

[27] A.C. Nascimento, R.B. Prudêncio, I.G.Costa, “A multiple kernel learning algorithm for drug–target interaction prediction,” BMC Bioinformatics, vol. 17, pp. 46 2016, DOI: 10.1186/s12859-016-0890-3.

[28] W. Lan, J. Wang, M. Li, J. Liu, Y. Li, et al., “Predicting drug–target interaction using positive - unlabeled learning”, Neurocomputing, vol. 206, pp. 50–57, 2016, DOI: 10.1016/j.neucom.2016.03.080.

[29] Z. Li, P. Han, Z.-H. You, X. Li, Y. Zhang, H. Yu, et al., “In silico prediction of drug-target interaction networks based on drug chemical structure and protein sequences,” Scientific Reports, vol. 7, no. 1, pp. 11174, 2017, DOI: 10.1038/s41598-017-10724-0.

[30] M. Ohue, T. Yamazaki, T. Ban, Y. Akiayama, “Link mining for kernel based compound–protein interaction predictions using a chemogenomics approach”, In: International Conference on Intelligent Computing, Springer, Cham, Switzerland, 2017, pp. 549–558, DOI: 10.1007/978-3-319-63312-1_48.

[31] J. Zhang, M. Zhu, P. Chen, B. Wang, “DrugRPE: random projection ensemble approach to drug–target interaction prediction”, Neurocomputing, vol. 228, pp. 256–262, 2017, DOI: 10.1016/j.neucom.2016.10.039.

[32] F. Rayhan, S. Ahmed, S. Shatabda, et al., “iDTI-ESBoost: identification of drug target interaction using evolutionary and structural features with boosting”, Scientific Reports, vol. 7, no. 1, pp. 17731, 2017, DOI:10.1038/s41598-017-18025-2.

[33] A. Sharma, R. Rani, “BE-DTI: ensemble framework for drug target interaction prediction using dimensionality reduction and active learning”, Computer Methods and Programs in Biomedicine, vol. 165, pp. 151–162, 2018, DOI:10.1016/j.cmpb.2018.08.011.

[34] F. Cheng, C. Liu, J. Jiang, et al., “Prediction of drug–target interactions and drug repositioning via network-based inference,” PLOS Computational Biology, vol. 8, no. 5, pp. e10025032012, 2012, DOI: 10.1371/journal.pcbi.1002503.

[35] X. Chen, M.-X. Liu, G.-Y. Yan, “Drug–Target Interaction prediction by random walk on the heterogeneous network,” Molecular Biosystems, vol. 8, no. 7, pp. 1970–1978, 2012, DOI: 10.1039/C2M00002D.

[36] H. Chen, Z. Zhang, “A semi-supervised method for drug–target interaction prediction with consistency in networks”, PloS One, vol. 8, no. 5, pp. e62975, 2013, DOI: 10.1371/joural.poe.0062975.

[37] L. Peng, B. Liao, W. Zhu, Z. Li, K. Li, “Predicting drug–target interactions with multi-information fusion”, IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 2, pp. 561–572, 2015, DOI: 10.1109/JBHI.2015.2513200.

[38] A. Seal, Y.-Y. Ahn, D.J. Wild, “Optimizing drug–target interaction prediction based on random walk on heterogeneous networks”, Journal of Cheminformatics, vol. 7, no. 1, pp. 40, 2015, DOI: 10.1186/s13321-015-0089-z.

[39] Y. Huang, L. Zhu, H. Tan, et al., "Predicting drug-target on heterogeneous network with co-rank," In: International Conference on Computer Engineering and Networks, Springer, Cham, Switzerland, 2018, pp. 571–581, DOI: 10.1007/978-3-030-14680-1_63.

[40] T. Ban, M. Ohue, Y. Akiyama, "NRLMFβ: beta-distribution rescored neighborhood regularized logistic matrix factorization for improving the performance of drug–target interaction prediction," Biochemistry and Biophysics Reports, vol. 18, pp. 100615, 2019, DOI: 10.1016/j.bbrep.2019.01.008.

[41] M. Wen, Z. Zhang, S. Niu, et al., "Deep-learning-based drug– target interaction prediction", Journal of Proteome Research, vol. 16, no. 4, pp. 1401–1409, 2017, DOI:1 0.1186/s12911-020-1052-0 .

[42] H. Öztürk, A. Özgür, E. Ozkirimli, "DeepDTA: deep drug–target binding affinity prediction", Bioinformatics, vol. 34, no. 17, pp. i821–i829, 2018, DOI: 10.1093/bioinformatics/bty593.

[43] L. Wang, Z.-H. You, X. Chen, et al, "A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network," Journal of Computational Biology, vol. 25, no. 3, pp. 361–373, 2018, DOI: 10.1089/cmb.2017.0135.

[44] J. You, R.D. McLeod, P. Hu, "Predicting drug–target interaction network using deep learning model," Computational Biology Chemistry, vol. 80, pp. 90–101, 2019, DOI: 10.1016/j.compbiolchem.2019.03.016.

[45] I. Lee, J. Keum, H. Nam, "DeepConv-DTI: prediction of drug-target interactions via deep learning with convolution on protein sequences", PLoS Computational Biology, vol. 15, no. 6, pp. e1007129, 2019, DOI: 10.1371/journal.pcbi.1007129.

[46] M. Kanehisa, M. Araki, S. Goto, et al., "KEGG for linking genomes to life and the environment," Nucleic Acids Research, vol. 36, pp. D480–484, 2007, DOI: 10.1093/nar/gkm882.

[47] M. Kanehisa, S. Goto, M. Hattori, M. Araki, M. Hirakawa, "From genomics to chemical genomics: new developments in KEGG," Nucleic Acids Research, vol. 34, pp. D354–D357, 2006, DOI: 10.1093/nar/gkj102.

[48] A. Gaulton, A. Hersey, M. Nowotka, et al., "The ChEMBL database in 2017," Nucleic Acids Research, vol. 45, no. D1, pp. D945–954, 2016, DOI: 10.1093/nar/gkw1074.

[49] J. Kringelum, S.K. Kjaerulff, S. Brunak, et al., "ChemProt-3.0: a global chemical biology diseases mapping", Database: the journal of biology databases and curation, vol. 2016 pp. bav123, 2016, DOI: 10.1093/database/bav123.

[50] A.H. Wagner, A.C. Coffman, B.J. Ainscough, et al, "DGIdb 2.0: mining clinically relevant drug–gene interactions", Nucleic Acids Research, vol. 44, no. D1, pp. D1036–1044, 2016, DOI: 10.1093/nar/gkv1165 .

[51] D.S. Wishart, Y.D. Feunang, A.C. Guo, et al., "Drugbank 5.0: a major update to the drugbank database for 2018", Nucleic Acids Research, vol. 46, no. D1, pp. D1074–1082, 2017, DOI: 10.1093/nar/gkx1037.

[52] M. Kanehisa, M. Furumichi, M. Tanabe, et al., "KEGG: new perspectives on genomes, pathways, diseases and drugs", Nucleic Acids Research, vol. 45, no. D1, pp. D353–361, 2016, DOI: 10.1093/nar/gkw1092.

[53] HMS_LINCS: LINCS Pilot Phase Joint Project: Sensitivity measures of six breast cancer cell lines to a library of small molecule kinase inhibitors (drug combination treatments). Dataset 2 of 2: Mean cell count and mean normalized growth rate inhibition values across technical replicates, 2016.

[54] J. Von Eichborn, M.S. Murgueitio, M. Dunkel, S. Koerner, P.E. Bourne, R. Preissner, "PROMISCUOUS: a database for network-based drug-repositioning", Nucleic Acids Research, vol. 36, Jan 2011, DOI:10.1093/nar/gkq1037.

[55] D. Szklarczyk, A. Santos, C. Von Mering, et al., "STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data", Nucleic Acids Research, vol. 44, no. D1, pp. D380–384, 2015, DOI: 10.1093/nar/gkv1277.

[56] S. Günther, M. Kuhn, M. Dunkel, et al., "Supertarget and matador: resources for exploring drug-target relationships", Nucleic Acids Research, vol. 36, pp. D919–922, 2008, DOI: 10.1093/nar/gkm862.

[57] X. Chen, Z.L. Ji, Y.Z. Chen, "TTD: therapeutic target database", Nucleic Acids Research, vol. 30, no. 1, pp: 412–415, 2002, DOI: 10.1093/nar/30.1.412.

[58] L. Jeske, S. Placzek, I. Schomburg, et al., "Brenda in 2019: a European ELIXIR core data resource", Nucleic Acids Research, vol. 47, no. D1, pp. D542–549, 2019, DOI: 10.1093/nar/gky1048.

[59] O. Ursu, J. Holmes, C.G. Bologa, et al., "DrugCentral 2018: an update," Nucleic Acids Research, vol. 47, no. D1, pp. D963–970, 2018, DOI: 10.1093/nar/gky963.

[60] C. Wang, G. Hu, K. Wang, et al., "PDID: database of molecular level putative protein–drug interactions in the structural human proteome", Bioinformatics, vol. 32, no. 4, pp: 579–586, 2016, DOI: 10.1093/bioinformatics/btv597.

[61] D.-T. Nguyen, S. Mathias, C. Bologa, et al., "Pharos: collating protein information to shed light on the druggable genome", Nucleic Acids Research, vol. 45, no. D1, pp. D995–D1002, 2017, DOI: 10.1093/nar/gkw1072.

[62] S. Kim, P.A. Thiessen, E.E. Bolton, et al., "PubChem substance and compound databases", Nucleic Acids Research, vol. 44, no. D1, pp. D1202–1213, 2016, DOI: 10.1093/nar/gkv951.

[63] V.B. Siramshetty, O.A. Eckert, B.-O. Gohlke, et al, "SuperDRUG2: a one stop resource for approved/marketed drugs", Nucleic Acids Research, vol. 46, no. D1, pp. D1137–1143, 2018, DOI: 10.1093/nar/gkx1088.

[64] H. Fang, Z. Su, Y. Wang, A. Miller, Z. Liu, P. C. Howard, W. Tong, & S. M. Lin, "Exploring the FDA adverse event reporting system to generate hypotheses for monitoring of disease characteristics", Clinical pharmacology and therapeutics, vol. 95, no. 5, pp. 496–498, 2014, DOI: 10.1038/clpt.2014.17.

[65] M. Kuhn, I. Letunic, L.J. Jensen, P. Bork, "The SIDER database of drugs and side effects," Nucleic Acid Research, vol. 44, no. D1, pp. D1075-1079, 2015, DOI: 10.1093/nar/gkv1075.

[66] A.J. Pawson, J.L. Sharman, H.E. Benson, et al., "The IUPHAR/BPS guide to pharmacology: an expert-driven knowledgebase of drug targets and their ligands," Nucleic Acids Research, vol. 42, no. D1, pp. D1098–1106, 2013, DOI: 10.1093/nar/gkt1143.

[67] R. Kumar, K. Chaudhary, S. Gupta, et al., "CancerDR: Cancer Drug Resistance Database", Scientific Reports, vol. 3, pp. 1445, 2013, DOI:10.1038/srep01445.

[68] M.K. Gilson, T. Liu, M. Baitaluk, et al., "BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology," Nucleic Acids Research, vol. 44, no. D1, pp. D1045–1053, 2016, DOI: 10.1093/nar/gkv1072.

[69] T. Sterling and J.J. Irwin, "ZINC- Ligand Discovery for Everyone," Journal of Chemical Information and modelling, vol. 55, no. 11, pp. 2324-2337, 2015, DOI: 10.1021/acs.jcim.5b00559.

[70] B.L. Roth, W.K. Kroeze, S. Patel, E. Lopez, "PDSP Ki -The Multiplicity of Serotonin Receptors: Uselessly diverse molecules or an embarrassment of riches?", The Neuroscientist, vol. 6, pp. 252–262, 2000, DOI:10.1177/107385840000600408.

[71] Y. Yamanishi, M. Kotera, M. Kanehisa, S. Goto, "Drug–target interaction prediction from chemical, genomic and pharmacological data in an integrated framework," Bioinformatics, vol. 26, no. 12, pp. i246–i254, 2010, DOI: 10.1093/bioinformatics/btq176.

[72] R. San-Miguel Carrasco, "Detection of Adverse Reaction to Drugs in Elderly Patients through Predictive Modeling", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 3, no. 6, pp. 52-56, 2016, DOI:10.9781/ijimai.2016.368.

A. Suruliandi

Dr. A. Suruliandi received the B.E., degree in electronics and communication engineering from the Coimbatore institute of Technology, Coimbatore, India, in 1987. He completed his M.E., degree in computer science and engineering from the Government College of Engineering, Tirunelveli, India, in 2000. And he pursued his Ph.D degree from Manonmaniam Sundaranar University, Tirunelveli, in 2009. He is currently working as a professor with the Department of Computer Science and Engineering, Manonmaniam Sundaranar University. He has more than 29 years of experience in teaching. He has been author of 50 articles in international journals, 23 articles in IEEE Xplore publications, 33 in national conferences, and 13 in international conferences. His interested research areas are remote sensing, image processing, and pattern recognition.

T. Idhaya

T. Idhaya received M.Sc., degree in Computer Science from St. Xavier's College (Autonomous), Tirunelveli, India, in 2016. She has completed her M.phil degree in Manonmaniam Sundaranar University, Tirunelveli, India, in 2017. She is currently pursuing her Ph.D degree in Manonmaniam Sundaranar University, Tirunelveli, India. Her area of interest is Image processing, Machine learning and Big data.

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. He published 75 papers in International Journals, 24 in International conferences and 12 in national conferences. Dr. Raja is an Associate Editor of the Journal of Circuits, Systems and Computers, Computing and Informatics, International Journal of Interactive Multimedia and Artificial Intelligence, Brazilian Archives of Biology and Technology, International Journal of Image and Graphics, and International Journal of Biometrics.

Brain Tumor Classification Using a Pre-Trained Auxiliary Classifying Style-Based Generative Adversarial Network

M. Akshay Kumaar[1], Duraimurugan Samiayya[2], Venkatesan Rajinikanth[3], P. M. Durai Raj Vincent[4], Seifedine Kadry[5,6,7]*

[1] BrainSightAI, Bangalore (India)
[2] St. Joseph's College of Engineering, Chennai (India)
[3] Department of Computer Science and Engineering, Division of Research and Innovation, Saveetha School of Engineering, SIMATS, Chennai 602105 (India)
[4] School of Information Technology and Engineering, Vellore Institute of Technology, Vellore (India)
[5] Department of Applied Data Science, Noroff University College, Kristiansand (Norway)
[6] Artificial Intelligence Research Center (AIRC), Ajman University, Ajman, 346 (United Arab Emirates)
[7] Department of Electrical and Computer Engineering, Lebanese American University, Byblos (Lebanon)

* Corresponding author. skadry@gmail.com

Received 21 March 2022 | Accepted 10 December 2022 | Early Access 9 February 2023

Abstract

Computer Vision's applications and their use cases in the medical field have grown vastly in the past decade. The algorithms involved in these critical applications have helped doctors and surgeons perform procedures on patients more precisely with minimal side effects. However, obtaining medical data for developing large-scale generalizable and intelligent algorithms is challenging in the real world as multiple socio-economic, administrative, and demographic factors impact it. Furthermore, training machine learning algorithms with a small amount of data can lead to less accuracy and performance bias, resulting in incorrect diagnosis and treatment, which can cause severe side effects or even casualties. Generative Adversarial Networks (GAN) have recently proven to be an effective data synthesis and augmentation technique for training deep learning-based image classifiers. This research proposes a novel approach that uses a Style-based Generative Adversarial Network for conditional synthesis and auxiliary classification of Brain Tumors by pre-training. The Discriminator of the pre-trained GAN is fine-tuned with extensive data augmentation techniques to improve the classification accuracy when the training data is small. The proposed method was validated with an open-source MRI dataset which consists of three types of tumors - Glioma, Meningioma, and Pituitary. The proposed system achieved 99.51% test accuracy, 99.52% precision score, and 99.50% recall score, significantly higher than other approaches. Since the framework can be made adaptive using transfer learning, this method also benefits new and small datasets of similar distributions.

Keywords

Generative Adversarial Image Classification, Image Synthesis, Magnetic Resonance Imaging, Medical Imaging, Networks.

DOI: 10.9781/ijimai.2023.02.008

I. Introduction

The healthcare industry has evolved tremendously because of the growing demands and population. It has become highly challenging for doctors to meet the velocity and volume of patients requiring diagnosis and treatments. The numerous realtime advances in Computer Vision and Artificial Intelligence in the medical field have been a game-changer in how fast and accurate most medical procedures are carried out. This has made doctors save time and establish treatments for patients early [1], [2]. While most of the algorithms in hand are not perfect, they are still helpful in real-time for gathering inferences at various stages of diagnoses and treatments.

Classification of Brain Tumors has been a difficult task. It is one of the most crucial steps and required information in diagnosis, presurgical planning, and treatment. While radiologists accurately identify and annotate these tumors for further steps, it is highly impossible for them to do the same at a large scale. Identifying the brain tumors and their type also depends on several scan parameters like the scan's modality, isotropy, magnetic field strength, and other acquisition parameters [3],

Please cite this article as: M. A. Kumaar, D. Samiayya, V. Rajinikanth, P. M. Durai Raj Vincent, S. Kadry, I. Ali, "Brain Tumor Classification Using a Pre-Trained Auxiliary Classifying Style-Based Generative Adversarial Network", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 6, pp. 101-111, 2024, http://dx.doi.org/10.9781/ijimai.2023.02.008

[4]. The scan acquisition also depends upon several socio-economic and demographic factors. Hospitals in a highly developed country may have 7 Tesla Magnetic Resonance Imaging machines. In contrast, countries with a not-so-similar economy and infrastructure may have only 1.5 Tesla Magnetic Resonance Imaging machines. These fields can vastly impact the manual and automated process of scan analysis during the diagnosis and planning of the treatment.

The recent development of several deep learning architectures like convolutional neural networks and transformers has helped solve and automate processes in several critical domains like healthcare, autonomous driving, and cybersecurity. Hence, deep learning can play a massive role in classifying brain tumors. The algorithms are also multi-disciplinary, enabling people to solve problems or invent in any field. However, the performance of these algorithms can be easily biased or degraded when it is not trained correctly with massive datasets using the appropriate methods. Obtaining medical datasets is challenging due to numerous legal procedures and factors like patient consent and the prevalence of disorder or disease. Geography also plays a vital role since people from different locations have different physical and mental attributes. This can cause the deep learning-based solutions can have less accuracy or bias based on the mentioned factors, leading to improper diagnosis and treatment that can result in side effects or even death. Several existing solutions have high false positive and negative rates for the same reasons. Furthermore, high accuracy or almost 100 % accuracy is expected in the healthcare industry since the algorithm's results are going to be used on patients for different purposes.

The groundbreaking invention of Generative Adversarial Networks (GAN) by Goodfellow et al. [5] has led to many creative applications and usage of image synthesis. GANs have also been popularly used as a data augmentation method that can improve the performance of image classifiers by oversampling. In this study, a framework that uses a Conditional Style-based Generative Adversarial Network with auxiliary classification [6], [7], [8], [9], [10] for pre-training the GAN and fine-tuning the Discriminator to improve the performance of auxiliary classification with extensive augmentation methods has been proposed. Fig. 5 displays a simplified architecture diagram of the system, giving an overview of how the proposed method works in real-time.

The proposed system uses an open-source dataset compiled by J. Cheng et al. [11] that comprises T1-weighted Magnetic Resonance Images with a magnetic field strength of 1.5 Tesla of many patients with three types of tumors: Glioma, Meningioma, and Pituitary. By training the proposed framework with extensive augmentation techniques during preprocessing, the system achieved a test accuracy of 99.51% and several other crucial validation metrics using the fine-tuned auxiliary classifying Discriminator of the pre-trained GAN. This method is advantageous when the availability of datasets is limited and can be fine-tuned using transfer learning for other datasets with similar distributions like scans from 3 Tesla or & Tesla machines.

This research article has been split into the following sections based on the experimental investigations:

1. The related research work that motivated the development and implementation of the proposed method.
2. Analysis of the dataset and the augmentation techniques was used as preprocessing for training the given framework.
3. A detailed description of the proposed architecture and its working with the training and validation strategies, hyperparameter settings, and optimization techniques that were followed.
4. The various evaluation metrics that were used to evaluate the model with the obtained results.
5. Discussion of the proposed system's results, advantages, and limitations.

II. Related Works

For this study, a deep analysis of the related research works and state-of-the-art methods for brain tumor classification was conducted, which helped in developing a novel and improved method for classifying brain tumors. Xiao et al. [12] proposed a method presenting the brain tumor classification with Dual Suppression Encoding and Factorized Bilinear Encoding with ResNet50 to differentiate minor features extracted from various MRI images belonging to different types of brain tumors. This achieved an excellent performance of 98.02% in classifying the exact features of brain tumors. While the feature engineering and the convolutional neural network architecture are amazing, the accuracy is comparatively lesser than some other methods.

Yerukalareddy et al. [6] introduced an intriguing method based on deep learning to classify brain tumors on MRI scans by pre-training MSGGAN to classify tumor types using the auxiliary block of the Discriminator. This approach obtained an accuracy of 98.57%, proving to be better than other methods. Our system uses a similar pre-training technique with significant changes in the neural network architecture, data preprocessing, and optimization strategies. Diaz-Pednas et al. [13] presented a multiscale approach using convolutional neural networks that used images at different resolutions and stages to learn feature representations to perform the classification. It was 97.3% successful in classifying meningioma, glioma, and pituitary types of tumors. Mohan Karnati et al. [14] presented a multi-scale deep convolutional neural network for detecting COVID-19 from X-rays with an accuracy >99%. This research works highlight the dynamicity and adaptability of neural networks to learn from any form of data.

Kumar R.L. et al. [15] have introduced a model that uses ResNet50 and global average pooling and acquired 97.08% and 97.48% efficient performance with and without data augmentation, respectively. While ResNet50 uses global average pooling by default, the authors have implemented a transfer learning-based approach to train the architecture to classify three types of brain tumors. Inspired by this approach, our framework was benchmarked with and without augmentation. Singh R. et al. [16] developed a Gabor-modulated convolutional filter-based tumor classification in the brain to classify Low-grade and High-grade Glioma. However, the number of network parameters is high, and it can only classify within the same tumor type. Likewise, Abd El Kader et al. [17] have derived a differential deep-CNN model to classify low-grade and high-grade glioma from brain MR images with an accuracy of 99.25%.

Kang et al. [18] presented a brain tumor classification method with an ensemble of features (DenseNet-169) using pre-trained convolutional neural networks, extracting the features of the tumor from MRI images using learned representations. They have achieved an accuracy of 93.72 % by using the CNN for feature extraction and Quasi-Support Vector Machine for classification. Alshayeji et al. [19] proposed an automatic classification of brain tumors by integrating two CNN structures and Bayesian optimization, resulting in higher performance with 97.37% correctness. Arbane et al. [20] used CNN architectures such as ResNet, Xception, and MobilNet-V2 based on transfer learning. They have compared these methods and concluded that MobilNet-V2 gave the best accuracy of 98.24% and an F1-score of 98.42%. Ayesha et al. [21] have invented a deep learning and improved particle swarm optimization-based algorithm to classify brain tumors using multiple MRI modalities with an accuracy of 99.9 %. Amjad Rehman et al. [22] proposed a 3D Convolutional Neural Network to extract brain tumors and employed a correlation-based feature selection to create a classification algorithm that achieved >92% on BraTS datasets. These research works helped in assessing the possible biases and challenges associated in training our algorithm.

Deepak S. et al. [23] adopted an approach of CNN with a support vector machine and attained a classification exactness of 95.82%. The system seems novel, but it is underperforming compared to other state-of-the-art methods. To classify the tumor, Singh R. et al. [24] have used a wavelet transform technique, which is fed to a kernel-based support vector machine. This algorithm yielded an accuracy of 98.87%, better than the deep neural-network approach. Ghassemi et al. [7] have used the GAN technique with a deep neural network pre-training process, resulting in high accuracies of 93.01% and 95.6% on the introduced and random splits, respectively, while classifying the meningioma, glioma, and pituitary tumors. The approach proposed in our study uses some motivations from this research work, and a robust framework was built like this.

Badza et al. [25] presented a new Convolutional Neural Network algorithm using a 10-fold cross-validation approach on the brain tumor databases. It produced 96.56% accuracy and claimed to have a decent generalization capability and execution speed. Deepak S. et al. [8] have used multi-scale gradient GAN to synthesize images of meningioma tumors. This approach has produced results close to the source dataset [26] but only for the Meningioma tumor type. Nonetheless, this method justifies the benefits of Generative Adversarial Networks in training image classifiers by using synthesized images for oversampling. Rehman A. et al. [27] have compared Convolutional Neural Network architectures like AlexNet, GoogLeNet, and VGGNet to extract distinguishable features and patterns from MRI images to obtain 98.69% accuracy using the VGG16 network. A. Seal et al. [28] proposed two probabilistic models using Logistic Regression (LR), Linear Discriminant Analysis (LDA) and a predictive model using Multilayer Perceptron (MLP) with a Fuzzy C-Means clustering algorithm for feature extraction of lesions in the human liver to predict whether a person has cancer in their liver or not. The MLP model achieved the lowest accuracy of 94.4 % when compared to the other models. This methodology would help extend the study's proposed method to accommodate for benign and malignant classification of the tumor in the future.

Our proposed method has used a similar approach to Oeldorf et al. [9], who have leveraged Conditional Style-based Generative Adversarial Networks to synthesize logos images. This is highly beneficial for synthesizing images based on desired class or condition and can be adapted to any appropriate dataset. Oeldorf et al. [9]'s paper was based on the Style-based Generative Adversarial Network proposed by Karras T. et al. [29], [30] from NVIDIA, improving the quality of the synthesized image using progressive growing of GANs, adaptive instance normalization, and style mixing using latent vectors. Karnewar et al. [31] have introduced a Multi-Scale Gradient Generative Adversarial Network for synthesizing high-resolution images, which uses images at different resolutions/scales at various stages to learn and synthesize images using unsupervised learning. Karnewar and team [32] have also improved their approach by generating synchronized multi-scale images using concatenation operation, limiting forced mixing regulation.

Sajjad M. et al. [30] have presented a Multi-Grade approach using a Deep Convolutional Neural Network, improving the correctness up to 94.58% using data augmentation and deep learning. This is useful in classifying the tumor grades, and the classification of tumor types of their method is sub-optimal. Seetha J. et al. [33] proposed an automated tumor detection mechanism using Convolutional Neural Networks with small kernels. Their approach attained an accuracy of 97.5% with minimum complication. Balasooriya M. et al. [34] developed a sophisticated deep learning method using CNN, performing with improved accuracy of 99.68%. Afshar P. et al. [35] have proposed to equip CapsNet incorporating raw and surrounding brain tissues, producing 90.89% accuracy. J. Cheng et al. [11] from Southern Medical University, Guangzhou, China open-sourced a brain tumor dataset containing T1- weighted contrast-enhanced images containing three types of tumors: glioma, meningioma, and pituitary. This dataset was used to train and benchmark the performance of the proposed method. Jun-Yan Zhu et al. [26] have proposed an approach for translating an image from a source domain to a target domain, and quantitative comparisons were demonstrated.

Goodfellow et al. [36] have designed a new way to synthesize non-linear probability distributions by using two neural network models that learn adversarially to improve each other;s performance with different goals called Generative Adversarial Networks. Kaiming et al. [28] presented a residual learning network called ResNet that uses skip connections at different convolutional blocks to improve feature learning and classification performance. Most of the related work and

TABLE I. Summary of Related Work

Authors	Method	Accuracy	Description	Limitations
Xiao et al. [12]	Dual Suppression Encoding and Factorized Bilinear Encoding	98.02%	A complex and robust feature extraction technique yielding good accuracy	The approach produced results very similar to other methods, and accuracy is sub-optimal
Yerukalareddy et al. [[6]	MSG-GAN pre-training and fine-tuning discriminator	98.57%	Advanced implicit feature learning by using GAN pre-training and fine-tuning the discriminator with augmented data	Generator and Discriminator requires up-sampled and down-sampled images at multiple stages to produce good results.
Diaz-Pednas et al. [13]	Multi-scale Convolutional Neural Networks	97.3%	Residual-like operation using source images at different resolutions at multiple stages to improve feature extraction	Requires sub-sampled images at multiple stages to produce good results.
Kumar R.L. et al.[15]	ResNet50	97.08%, 97.48%	Transfer learning with ResNet50 architecture on augmented and non-augmented data	ResNet50's pre-trained weights are of ImageNet dataset which does not have learned features from MRI scans thus having decent results.
Singh R. et al. [24]	Wavelet-based transformation with a kernel-based Support Vector Machine	93.72%	A different feature extraction technique based on image processing having good results using SVM.	Convolutional Neural Networks have proven to be better at learning kernels dynamically but the proposed SVM has achieved sub-par results.
Ghassemi et al. [7]	ACGAN based pre-training and fine-tuning of the discriminator	95.6%	Advanced implicit feature learning by using GAN pre-training and fine-tuning the discriminator with data of different splits.	The GAN architecture is based on DCGAN which upon training can be unstable and can cause mode-collapse with limited samples.

work done so far was using Convolutional Neural Networks using the figshare dataset provided by J. Cheng [11]. Despite multiple similar approaches with different training and hyper-parameter optimization-based strategies, most of the works' accuracies are not up to the mark. The GAN pre-training-based approaches [6], [7] have achieved almost perfect classification results without using the images synthesized for oversampling and training different network architectures. The literature survey of related works has been summarized in Table I.

These research works helped in developing a sophisticated system that can classify three types of tumors using an auxiliary classifying style-based generative adversarial network. While most of the related works have used deep learning-based approaches, the usage of different architectures, datasets, hyperparameters, preprocessing, and training strategies have led to models with variable biases and high false positive/negative rates in different cases. In Section III, the research methodology has been explained in detail, giving an overview of the dataset, preprocessing techniques, and proposed system implementation. The results obtained from the conducted experiments were validated using different strategies and evaluation metrics, which have been discussed in Section IV. The advantages, limitations, and analysis against several state-of-the-art methods have been briefly discussed in Section V. Finally, the proposed research and the future works to overcome the drawbacks of the proposed method have been summarized in Section VI.

III. Methodology

The proposed method uses an Auxiliary Conditional Style-based Generative Adversarial Network for pre-training and usage of the pre-trained Discriminator of the GAN by fine-tuning for classifying Glioma, Meningioma, and Pituitary tumors from a given MR Image. The experiments were conducted on a system with Ubuntu OS with 54 GB RAM and eight 16 GB NVIDIA V100 graphic processing units for faster training using distributed computing. Python was used to develop experiments with the help of libraries like PyTorch, NumPy, Matplotlib, Seaborn, Pandas, and Scikit-Learn. The methodology and experiments performed on the dataset, preprocessing steps, and neural network architectures have been explained in detail in the following sub-sections.

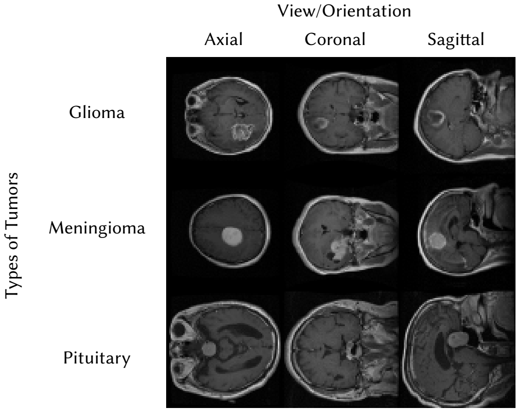


Fig. 1. T1-Weighted MR Images of tumor types in different views.

A. Dataset

The system was trained and benchmarked using a dataset open-sourced on Figshare in 2017 by J. Cheng et al. [11] containing T1-weighted Magnetic Resonance Images of Glioma, Meningioma, and Pituitary tumors. There are 3064 contrast-enhanced images presented as 2D MRI scans from 233 patients. The images appear to be of 1.5 Tesla Magnetic Field Strength and are of 512x512 resolution in coronal, axial, and sagittal views, as seen in Fig. 1. However, the number of images across all three modalities is low and can lead to a biased or less accurate classification rate upon training deep learning or machine learning algorithms. Table II describes the dataset in detail with supplemental information.

B. Preprocessing & Preparation

The dataset was downloaded from figshare as uploaded by J. Cheng et al. [11]. The MR Image slices were extracted from big data file format (.h5) and saved as png images under the corresponding folder as their class name, denoting the type of brain tumor. Convolutional neural networks perform better when the resolution of the image is

TABLE II. T1-Weighted Brain Tumor MRI Dataset Details

Type of Brain Tumor	Number of Patients	Number of MR Image slices	View/Orientation			Ground Truths	
			Axial View	Coronal View	Sagittal View	Tumor Labels	Tumor Masks
Glioma	89	1426	✓	✓	✓	✓	✓
Meningioma	82	708	✓	✓	✓	✓	✓
Pituitary	62	930	✓	✓	✓	✓	✓
Total	233	3064					

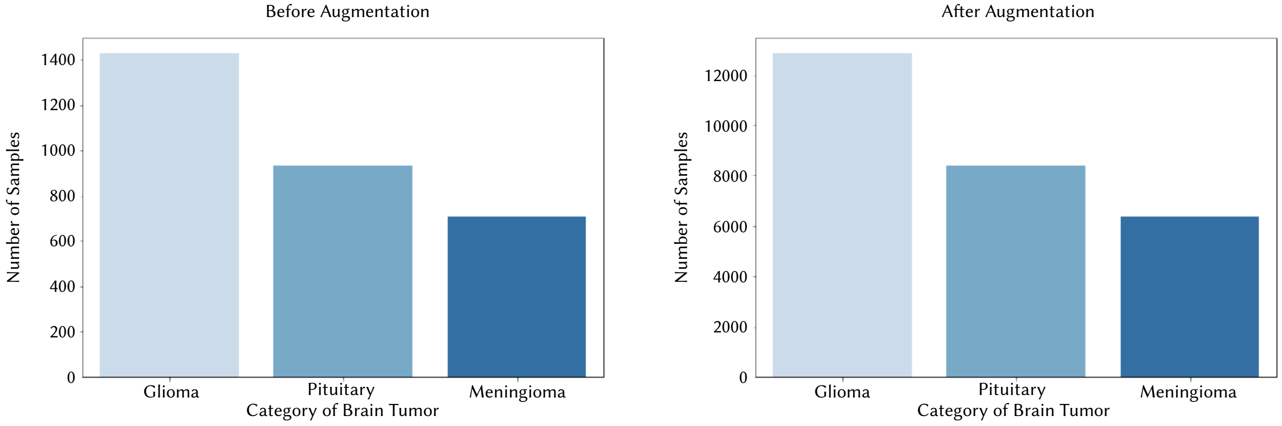


Fig. 2. Count Plot of samples before augmentation (up) and after augmentation (down).

higher. However, most MR Images are obtained in 256x256 resolution as most scanners have 256 as their frequency and Field-of-View (FOV) parameters since it takes less time for acquiring a T1-weighted MRI. MR Image scanners are also susceptible to noise during scan acquisition, and the noise produced is mostly Gaussian or non-linear. A matrix is randomly sampled from Gaussian distribution as described by (1) and (2) and is added to the input image to introduce noise in the source MR Image.

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{1}$$

$$f(x) = \begin{bmatrix} P(z_{11}) & P(z_{12}) & P(z_{13}) & \dots & P(z_{1n}) \\ P(z_{21}) & P(z_{22}) & P(z_{23}) & \dots & P(z_{2n}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ P(z_{d1}) & P(z_{d2}) & P(z_{d3}) & \dots & P(z_{dn}) \end{bmatrix} + \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & x_{d3} & \cdots & x_{dn} \end{bmatrix} \tag{2}$$

Histogram equalization was performed on the data to standardize the distribution and equalize intensity values, removing any bias fields. Finally, angular augmentations were performed on the image to get resultant images in 45, 90, 120, 150, 180, 270, 300, 330 degrees. Even though the scans in the real world would be in -90 to 90, -180 to 180 degrees in axial, coronal, and sagittal views, this type of augmentation would help the model learn the kind of tumors at different places and orientations. The labels were then one-hot encoded to create sparse tensors representing the tumor type. Fig. 2 depicts the dataset's sample distribution pre and post data augmentation.

These were the augmentation techniques performed for preprocessing the MR Images in general. However, two types of preprocessing with different augmentation techniques were performed for training the GAN and then fine-tuning the Discriminator of the pre-trained GAN:

1. Strategy 1: The augmentation technique used for training the GAN is resizing the input image to 256x256 resolution, applying random center cropping, random translation of the image towards left or right randomly, and histogram equalization. This augmentation resulted in 3064 images with some source images randomly cropped and translated.
2. Strategy 2: The augmentation techniques followed for fine-tuning the discriminator model were resizing the input image to 256x256 resolution, applying random center cropping, random Gaussian noise, histogram equalization, and angular augmentation. This augmentation resulted in 27576 images that contained source images, and the augmentation applied images.

The preprocessing can be done on the fly based on any given batch size using PyTorch Datasets and DataLoaders to save memory and GPU usage. A stratified 5-fold cross-validation was performed on the dataset, which splits the given dataset into five random subsets while preserving the ratio of the number of samples per class in each subset where the algorithm is trained by combining four of the subsets and evaluating it against the remaining set in all permutations and combinations. These approaches were used for pre-training the GAN and fine-tuning the Discriminator using corresponding augmentation methods.

C. Algorithm

The suggested method is based on a conditional Style-based Generative Adversarial Network with an auxiliary classification block that performs the tumor classification. It comprises two significant portions: pre-training the GAN and fine-tuning the pre-trained discriminator network with heavy data augmentation. The style-based generator is a modified architecture proposed by Karras T. et al. [10] with conditional input support to help the generator learn and produce distributions based on specified input classes like the method proposed by Oeldorf et al. [9]. The generator of the GAN uses convolutional blocks of decreasing filter size (256 -> 128 -> 64 -> 32) with a 3x3 kernel followed by adaptive instance normalization and leaky relu activation with bilinear upsampling. The Discriminator of the GAN is a simple convolutional neural network comprising of convolutional layers with growing filter size (16 -> 32 -> 64 -> 128 -> 256) followed by leaky relu (4) and max-pooling, with two final output layers: adversarial fully-connected layer (5) for image legitimacy prediction and an auxiliary

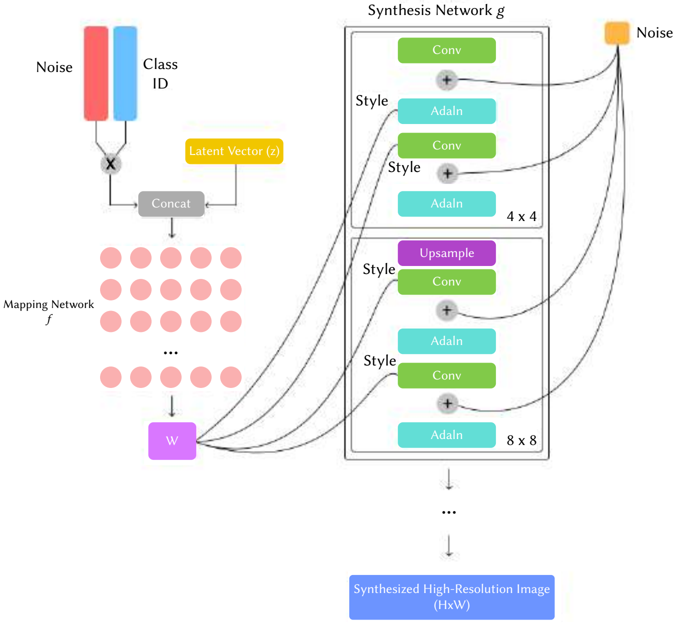


Fig. 3. Conditional Generator for the Style-based Generator Adversarial Network.

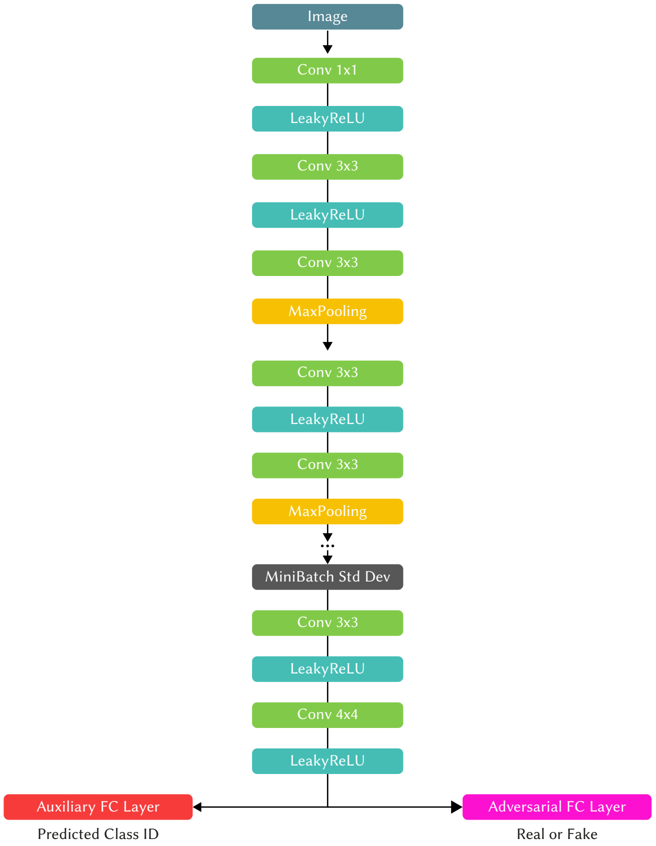


Fig. 4. Discriminator of the Style-based Generative Adversarial network with an auxiliary output block for tumor classification.

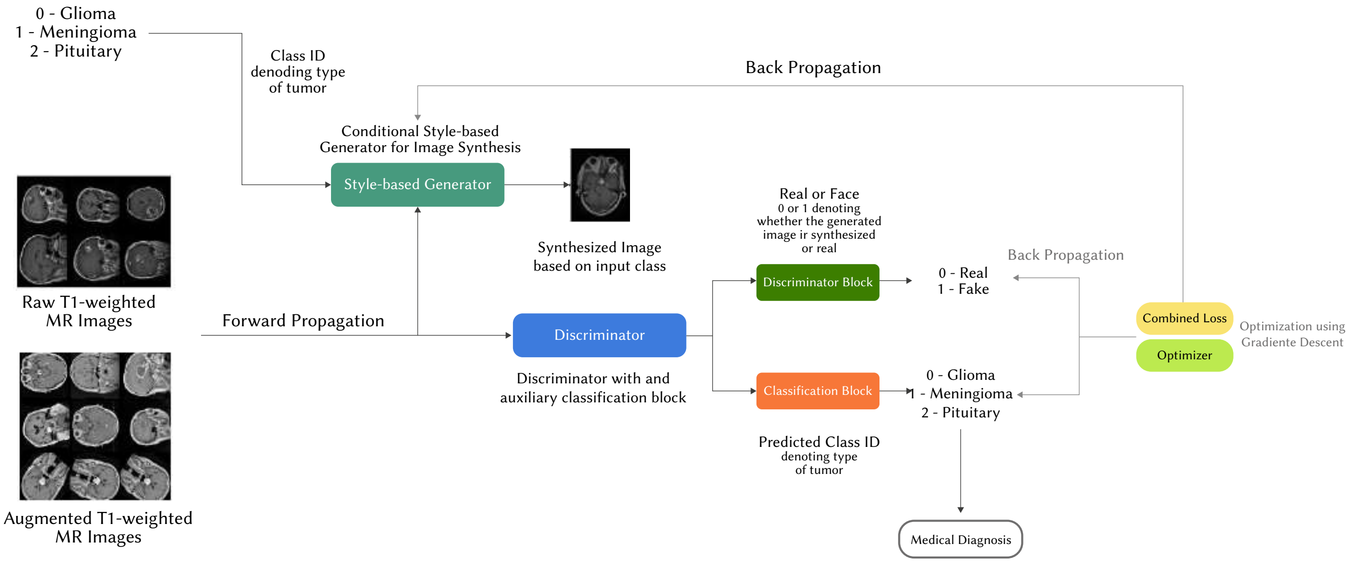


Fig. 5. System Architecture Diagram.

fully-connected layer for tumor classification. The hyperparameters for the convolutional blocks in both Generator and Discriminator are set to grow progressively as suggested by Karras T. et al. [37] to have better learning fidelity. Furthermore, the convolutional layers' filter size can grow from 16 to 1024, for a target image resolution of 1024x1024. A detailed architecture diagram of the described generator and discriminator neural networks is displayed in Fig. 3 and Fig. 4.

The process of the Generative Adversarial Network can be formulated as (3). Unlike Yerukulareddy et al. [6], the proposed generator uses a concatenated noise vector containing class representations and the latent vector as an input instead of classwise embedding, represented in Fig. 3. This makes the input latent vectors less sparse and leads to better feature identification and learning since the input has class details with a random distribution to generate the images.

$$\min_{G}\max_{D}\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log\ D(x)]+\mathbb{E}_{z\sim p_{\text{generated}}(z)}[1-\log\ D(G(z))] \tag{3}$$

$$R(z)=\begin{Bmatrix} z & z>0 \\ \alpha z & z<=0 \end{Bmatrix} \tag{4}$$

$$\hat{y}=W^{T}.X+b \tag{5}$$

The concatenated input latent vectors are passed through a fully connected network, called the mapping network, which maps styles or feature representations to the synthesis network for conditional image generation. The progressive growth or upsampling of the images combined with the mapping network, adaptive instance normalization (6) [38], and latent vectors with random and class representations helps the generator to learn classwise features adversarially with feedback received from the Discriminator regarding the image legitimacy and correctness of the tumor image generated.

$$\text{AdaIN}\ (\boldsymbol{x}_i,\boldsymbol{y}_i)=\boldsymbol{y}_{s,i}\left(\frac{\boldsymbol{x}_i-\mu(\boldsymbol{x}_i)}{\sigma(\boldsymbol{x}_i)}\right)+\boldsymbol{y}_{b,i} \tag{6}$$

The Discriminator has two objectives: the adversarial fully connected block predicts image legitimacy, and the auxiliary fully connected block predicts tumor type. The discriminator is a progressively growing convolutional neural network with a two-way output channel, as seen in Fig. 4. The adversarial fully connected block is activated with a sigmoid (7) to output probability-like values within the range of 0 to 1. Softmax (8) was used to activate the auxiliary fully connected block to output a vector of 3 probability-like values representing the type of tumor where the index with maximum probability can be mapped to the tumor's name.

$$\sigma(z)=\frac{1}{1+e^{-z}} \tag{7}$$

$$\sigma(z)_j=\frac{e^{z_j}}{\sum_{k=1}^{K}e^{z_j}} \tag{8}$$

Binary Cross-entropy (9) is used as the loss function for the adversarial outputs, and categorical cross-entropy (10) is used as the loss function for the auxiliary outputs. These two losses are averaged and used for backpropagation with gradient penalty, similar to WGAN-GP [39]. The GAN's discriminator loss and the combined loss function with gradient penalty are represented by (11) and (12).

$$\text{BinaryCrossentropy}=\text{bee}\ (y,p)=-(y\log\ (p)+(1-y)\log\ (1-p)) \tag{9}$$

$$\text{CategoricalCrossentropy}=cce(y,p)=-\sum_{c=1}^{M}y_{o,c}\log\ (p_{o,c}) \tag{10}$$

$$\mathit{Discriminator\,Loss}=\nabla_{\theta D}\left[f_{\theta D}\left(\left(bce(y,\hat{y})+cce(y,\hat{y})\right)/2\right)-\sum_{c=1}^{M}y_{o,c}log(p_{o,c})\right] \tag{11}$$

$$\text{Combined Loss}\ =\ \text{Discriminator Loss}\ +\lambda\left[\left(\|\nabla_{\hat{x}}D(\hat{y},y)\|_2-1\right)^2\right] \tag{12}$$

Pre-training of the GAN was done with a latent size of 512 for 1000 epochs with a batch size of 128 at a learning rate of 0.005 using the first augmentation strategy mentioned in section 3.2. Once the GAN is trained, the pre-trained Discriminator's weights were frozen for all the layers except the auxiliary classification layer. This Discriminator was then fine-tuned with a batch size of 64 at a learning rate of 1e-4 with the second augmentation strategy given under section 3.2 and without any augmentation using categorical cross-entropy (10) as loss function. Later, Adam optimizer [40] for optimization while training the GAN and fine-tuning the pre-trained discriminator network. Table 3 presents the optimization parameters for training the GAN and fine-tuning the Discriminator.

The entire system's architecture in a real time setting can be seen in Fig. 5. In the upcoming sections, the evaluation metrics used for evaluating the proposed classifier and the achieved results are discussed, with a comparison of it against the other existing state-of-the-art methods.

TABLE III. Parameters Used for Optimizing the Network During Pre-Training and Fine-Tuning

Hyper-parameters	GAN Pre-training	Discriminator Fine-tuning
Optimizer	Adam	Adam
Loss Function	Combined Loss (9)	Categorical Crossentropy (7)
Latent Size	512	-
Epochs	1000	10
Batch Size	128	64
Augmentation Strategy	Strategy 1 given in section 3.2	Resizing images to 256x256, no augmentation strategy Strategy 2 given in section 3.2
Learning Rate	0.005	1e-4

IV. Results

The benchmarking experiments were done using the Discriminator network of the pre-trained GAN framework on the test samples of 5-fold cross-validation sets with and without augmentation, as mentioned in section III.B. A brief description of the evaluation metrics used is discussed in section IV.I, and the obtained results using the experimental setup are displayed in section IV.II.

A. Evaluation Metrics

The following evaluation metrics have been used for analyzing the results of the proposed method to understand the algorithm's performance and limitations:

As seen in Table IV, the confusion matrix has been used to understand the model's classification performance and derive other metrics that can give insights into bias and limitations of the algorithm in place.

TABLE IV. Typical Confusion Matrix

	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

The correctness of the model can be defined by accuracy, which tells how right the model has done the classification.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

The ratio of true positives to the sum of true positives and false positives is precision. The ratio of true positives and the sum of true positives and false negatives are known as recall. The balance between these scores is known as the F1 score. These scores help us understand the true classification rate of the classifier in depth.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{14}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{15}$$

$$F1 = \frac{2 * \text{ Precision } * \text{ Recall}}{\text{Precision } + \text{ Recall}} = \frac{2 * TP}{2 * TP + FP + FN} \tag{16}$$

The rate of true positives against false negatives is described by a Receiver Operating Characteristic (ROC) Curve, which can tell how well a classifier is good at producing true positives. A classification algorithm performs better when its ROC Area Under the Curve (AUC) score is higher.

$$\text{Sensitivity} = \text{Recall} = \frac{TP}{TP + FN} \tag{17}$$

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{18}$$

$$AUC = \frac{\text{Specificity } + \text{ Sesnsitivity}}{2} \tag{19}$$

B. Model Evaluation

The discriminator network of the pre-trained GAN was fine-tuned on the 5-fold cross-validation sets with and without augmentations using different seed values for sampling and internal shuffling. The predictions were converted to sparse tensors containing one at indices with maximum confidence/probability values in the output tensor that denote the type of tumor using the argmax operation (20).

$$\text{result} = \arg \underset{\theta}{\text{m}}\ g(x) \tag{20}$$

Tables V and VI display the model's performance data on the mentioned test sets with accuracy, precision, recall, and F1 scores as evaluation metrics.

TABLE V. Fine-tuned Discriminator's Performance on Non-augmented Test Sets of All the Five Folds

Fold/Evaluation Metrics	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Fold 1	99.80	99.68	99.42	99.55
Fold 2	**99.83**	**99.88**	**99.76**	**99.82**
Fold 3	99.45	99.56	99.40	99.48
Fold 4	98.98	99.03	98.88	98.95
Fold 5	99.63	99.72	99.54	99.63
Mean	99.53	99.57	99.40	99.48

Table VI shows that the fine-tuned model obtained a whopping accuracy of 99.83% on the test set of non-augmented Fold 2 and 99.51% accuracy on the test set of augmented Fold 3. The mean accuracy of the model on non-augmented and augmented sets is around 99.53% and 99.21%. The results of the fine-tuned discriminator network of pre-trained GAN on the test sets of the best-performing fold with and without augmentations are highlighted in Tables V and VI.

TABLE VI. Fine-tuned Discriminator's Performance on Augmented Test Sets of All the Five Folds

Fold/Evaluation Metrics	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Fold 1	98.78	98.49	98.56	98.52
Fold 2	99.1	98.95	99.02	98.98
Fold 3	**99.51**	**99.52**	**99.50**	**99.51**
Fold 4	99.42	99.44	98.39	98.91
Fold 5	99.27	99.26	99.23	99.24
Mean	99.216	99.13	98.94	99.03

Fig. 6 and Fig. 7 are the confusion matrix and Receiver Operating Characteristic Curve for all the classes, obtained using the second subset of non-augmented 5-fold stratified cross-validation set. Only one image was misclassified in the test set, and the ROC-AUC scores are higher, suggesting that the model is great at classifying between the three types of tumors. Fig. 8 displays the classwise ROC curves with their corresponding AUC scores.

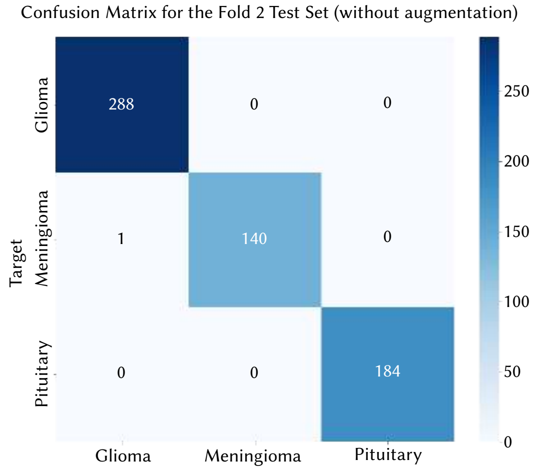


Fig. 6. Confusion Matrix for the non-augmented test set (Fold 2).

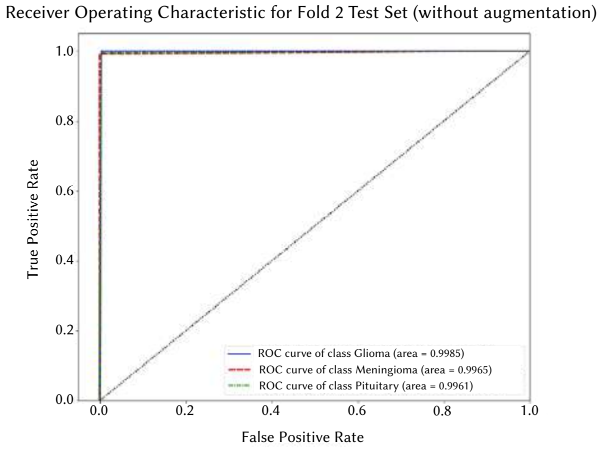


Fig. 7. Receiver Operating Characteristic (ROC) curve for the non-augmented test set (Fold 2).

The confusion matrix and Receiver Operating Characteristic Curve for all classes are shown in Fig. 9 and Fig. 10, obtained using the third fold's test set of the augmented 5-fold stratified cross-validation set. Out of the 5514 images in the test set, there were only 27 that were incorrectly classified. The model's ROC-AUC scores are also high, indicating that the model is very good at classifying the three types of tumors. The classwise ROC curves with their corresponding AUC scores are shown in Fig. 11.

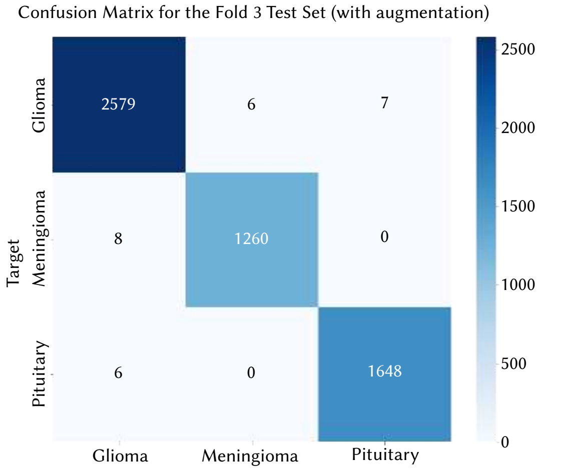


Fig. 9. Confusion Matrix for the augmented test set (Fold 3).

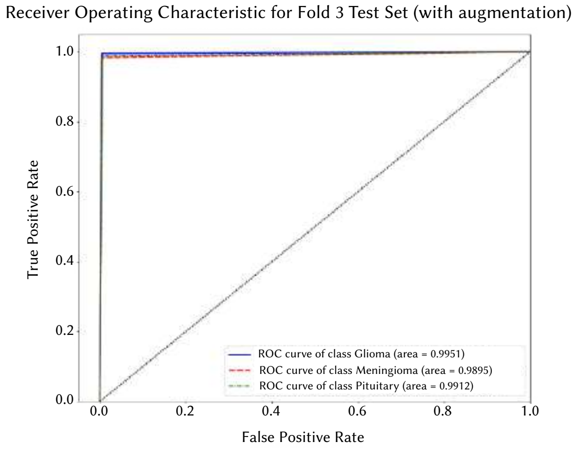


Fig. 10. Receiver Operating Characteristic (ROC) curve for the augmented test set (Fold 3).

V. Discussion

This section evaluates the model's performance against some state-of-the-art brain tumor classification methods. The neural network architectures of the other methodologies were implemented to compare the results with the proposed method. The experimental investigations showed that the fine-tuned discriminator network has over 99.5% in successfully classifying the three tumor types. Table

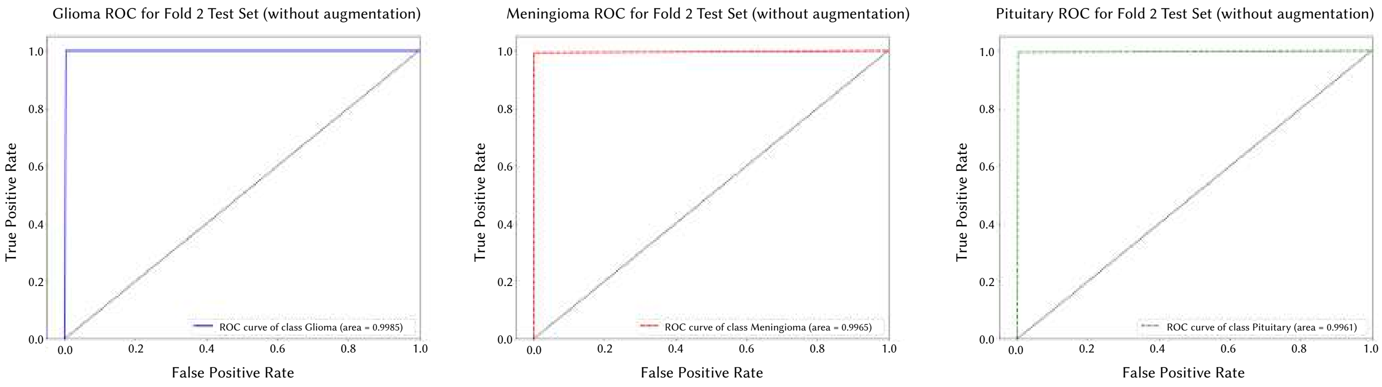


Fig. 8. Classwise ROC curves with AUC values for the non-augmented test set.

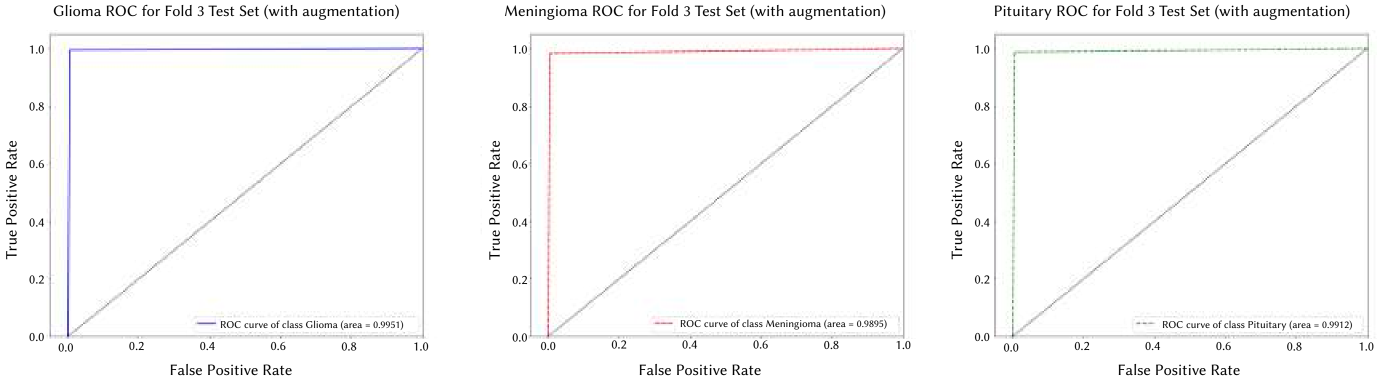


Fig. 11. Classwise ROC curves with AUC values for the augmented test set.

VII displays some of the evaluation metrics obtained by the proposed method against other existing methods.

From Table VII, it can be said that the proposed method has achieved the best metrics. The pre-training mechanism with minimal augmentation methods, as mentioned in section 3.2, used while training the conditional style-based GAN with auxiliary classification helps the discriminator network learn even the most minor features to identify it is real or fake as well as the type of tumor. This also increases the generator's performance adversarial to synthesize images of tumors. The framework learns to synthesize images and predict image legitimacy and the type of tumor simultaneously while training it for auxiliary classification and image generation. When the pre-trained Discriminator is fine-tuned, all the convolutional layers are frozen, and only the auxiliary classification block is trained to improve the classification performance with extensive augmentation. This makes the fully connected layer use the learned features from the convolutional blocks that act as feature extractors and adapt to the target output classes with features learned during the pre-training for image generation and legitimacy prediction. This also indicates that the model is adaptive and highly beneficial in transfer learning to apply it for datasets of similar distribution like brain MRI or other MRI scans from 3 Tesla machines or 7 Tesla machines. The entire architecture can be retrained on any dataset since deep learning algorithms are naturally adaptive to multifaceted applications.

TABLE VII. Comparison of the Proposed Method Against Existing Approaches

Method	Accuracy (%)	Precision (%)	Recall (%)
Proposed Method - with the augmented test set	99.51	99.52	99.50
Proposed Method - with the raw test set	99.83	99.88	99.76
Synergy Factorized Bilinear Network with a Dual Suppression [12]	97.96	97.43	97.67
MSG-GAN Pre-trained model [6]	98.62	98.65	98.71
Multiscale Convolutional Neural Network [13]	97.30	97.42	97.35
ResNet50 with angular augmentation [15]	97.49	97.51	97.54
AC-GAN Pre-trained model [7]	95.60	95.29	95.10

Since the model is pre-trained and fine-tuned on a T1-weighted MRI dataset, it would not perform well for other scan modalities like T2, Flair, Contrast, and functional MRI and scans with different magnetic field strengths. Likewise, the model has not been experimented with other hyperparameters for a different target image resolution (example: 1024x1024, 512x512). However, the framework can be retrained or fine-tuned using an appropriate dataset to improve its performance for other scans or data distributions using transfer learning. The style-based generator synthesizes good images, sometimes producing subpar results. This kind of framework trains datasets with implicit oversampling and improves the classifier's performance. However, the image synthesis of brain tumors may not have any real-world usage apart from using them for oversampling while training deep learning models. Also, the generated images have to be clinically validated by radiologists to know how accurately the GAN can synthesize images of the tumors and how useful it is for doctors. Fig. 12 and Fig. 13 show some of the generated images that were good and those that were bad.

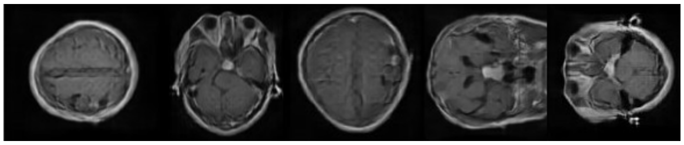


Fig. 12. Generated samples by the GAN.

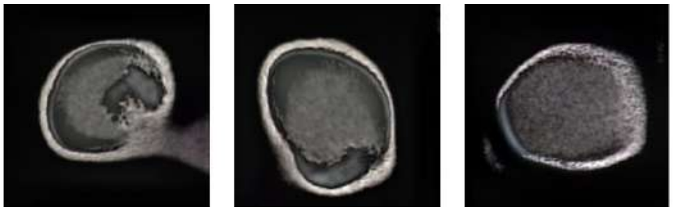


Fig. 13. Badly generated samples by the GAN.

VI. Conclusion

The proposed approach for the multi-class classification of brain tumors uses pre-training on an auxiliary classifying Style-based Generative Adversarial Network [9, 29, 10] to classify Glioma, Meningioma, and Pituitary tumor types using T1-weighted MR Images. The framework has two major processes: pre-training the conditional style-based GAN and fine-tuning the pre-trained Discriminator with extensive data augmentation to improve the classification performance. Pre-training the GAN with different augmentation strategies helps the algorithm learn feature representations from the data in a semi-supervised approach while also enabling the Discriminator to predict the legitimacy of the image and type of tumor present in the images. The method has achieved an accuracy of 99.51% on the augmented test set 99.83% on the raw test set, which is comparatively better than the other proposed approaches. The system is also sound when the availability of data is less. However, the model must be trained with data from multiple distributions containing different modalities and from machines of different magnetic field strengths like 1.5 Tesla, 3 Tesla and 7 Tesla machines to achieve better generalizability and classification performance, which can be done using transfer learning. To overcome these disadvantages, we plan to introduce few-shot learning or self-supervised architectures with adversarial pre-training and augmentation on a diverse multi-modal dataset to achieve the highest possible performance, fairness, and robustness for classifying brain tumors as well as classifying whether the tumor is benign or malignant [28].

References

[1] M. Sharif, J. Li, M. Khan, S. Kadry, and U. Tariq, "M3BTCNet: multi model brain tumor classification using metaheuristic deep neural network features optimization," *Neural Computing and Applications*, vol. 36, pp. 95-110, 2018, doi: 10.1007/s00521-022-07204-6.

[2] M. Nawaz, T. Nazir, M. Masood, A. Mehmood, R. Mahum, M.A. Khan, S. Kadry, O. Thinnukool, "Analysis of Brain MRI Images Using Improved CornerNet Approach," *Diagnostics*, vol. 11, no. 10, pp. 1-18, 2022, doi: 10.3390/diagnostics11101856.

[3] A. Aziz, M. Attique, U. Tariq, Y. Nam, M. Nazir, C. Jeong, R. R. Mostafa, R. H. Sakr, "An Ensemble of Optimal Deep Learning Features for Brain Tumor Classification," *Computers, Materials and Continua*, vol. 69, no. 2, pp. 2653-2670, 2021, doi: 10.32604/cmc.2021.018606.

[4] M.I. Sharif, M.A. Khan, M. Alhussein, K. Aurangzeb, M. Raza, "A decision support system for multimodal brain tumor classification using deep learning," *Complex & Intelligent Systems*, vol. 8, 2022, doi: 10.1007/s40747-021-00321-0.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[6] D. R. Yerukalareddy and E. Pavlovskiy, "Brain Tumor Classification based on MR Images using GAN as a Pre-Trained Model," in *Proceedings of the 2021 IEEE Ural-Siberian Conference on Computational Technologies in Cognitive Science, Genomics and Biomedicine (CSGB)*, pp. 380-384, 2021, doi: 10.1109/CSGB53040.2021.9496036.

[7] N. Ghassemi, A. Shoeibi, and M. Rouhani, "Deep neural network with generative adversarial networks pre-training for brain tumor classification based on MR images," *Biomedical Signal Processing and Control*, vol. 57, pp. 101678, 2020, doi: 10.1016/j.bspc.2019.101678.

[8] S. Deepak and P. M. Ameer, "MSG-GAN Based Synthesis of Brain MRI with Meningioma for Data Augmentation," in *Proceedings of the 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pp. 1-6, 2020, doi: 10.1109/CONECCT50063.2020.9198672.

[9] C. Oeldorf and G. Spanakis, "LoGANv2: Conditional Style-Based Logo Generation with Generative Adversarial Networks," pp. 462-468, 2019, doi: https://arxiv.org/abs/1909.09974.

[10] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396-4405, 2019, doi: 10.1109/CVPR.2019.00453.

[11] J. Cheng, "Brain Tumor Dataset 2017," 2017. [Online]. Available: https://doi.org/10.6084/m9.figshare.1512427.v5.

[12] G. Xiao et al., "Synergy Factorized Bilinear Network with a Dual Suppression Strategy for Brain Tumor Classification in MRI," *Micromachines*, vol. 13, no. 1, pp. 15, 2022, doi: 10.3390/mi13010015.

[13] F. Díaz-Pernas, M. Martínez Zarzuela, M. Antón-Rodríguez, and D. González-Ortega, "A Deep Learning Approach for Brain Tumor Classification and Segmentation Using a Multiscale Convolutional Neural Network," *Healthcare*, vol. 9, pp. 153, 2021, doi: 10.3390/healthcare9020153.

[14] M. Karnati, A. Seal, G. Sahu, and A. Yazidi, "A novel multiscale-based deep convolutional neural network for detecting COVID-19 from X-rays," *Applied Soft Computing*, vol. 125, 2022, doi: 10.1016/j.asoc.2022.109109.

[15] R. L. Kumar, J. Kakarla, B.V. Isunuri, M. Singh, "Multi-class brain tumor classification using residual network and global average pooling," *Multimedia Tools and Applications*, vol. 80, pp. 13429–13438, 2021, doi: 10.1007/s11042-020-10335-4.

[16] R. Singh, A. Goel, and D. K. Raghuvanshi, "Computer-aided diagnostic network for brain tumor classification employing modulated Gabor filter banks," *The Visual Computer*, vol. 37, pp. 2157–2171, 2020, doi: 10.1007/s00371-020-01977-4.

[17] I. Abd El Kader, X. Guizhi, Z. Shuai, S. Saminu, I. Javaid, and I. S. Ahmad, "Differential Deep Convolutional Neural Network Model for Brain Tumor Classification," *Brain Sciences*, vol. 11, no. 3, pp. 352, 2021, doi: 10.3390/brainsci11030352.

[18] J. Kang, Z. Ullah, and J. Gwak, "MRI-Based Brain Tumor Classification Using Ensemble of Deep Features and Machine Learning Classifiers," *Sensors*, vol. 21, no. 6, pp. 2222, 2021, doi: 10.3390/s21062222.

[19] M. Alshayeji, J. Al-Buloushi, A. Ashkanani, S. Abed, "Enhanced brain tumor classification using an optimized multi-layered convolutional neural network architecture," *Multimedia Tools and Applications*, vol. 80, pp. 28897–28917, 2021, doi: 10.1007/s11042-021-10927-8.

[20] M. Arbane, R. Benlamri, Y. Brik, and M. Djerioui, "Transfer Learning for Automatic Brain Tumor Classification Using MRI Images," in *Proceedings of the 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*, 2021, pp. 210-214, doi: 10.1109/CSGB53040.2021.9496036.

[21] A. B. T. Tahir, M.A. Khan, M. Alhaisoni, J. Ali Khan, Y. Nam, S. Wang, K. Javed, "Deep Learning and Improved Particle Swarm Optimization Based Multimodal Brain Tumor Classification," *Computers, Materials and Continua*, vol. 68, no. 1, pp. 1099-1116, 2021, doi: 10.32604/cmc.2021.015154.

[22] A. Rehman, M. Khan, T. Saba, Z. Mehmood, U. Tariq, and N. Ayesha, "Microscopic Brain Tumor Detection and Classification using 3D CNN and Feature Selection Architecture," *Microscopy Research and Technique*, vol. 84, pp. 133-149, 2020, doi: 10.1002/jemt.23597.

[23] S. Deepak and P. M. Ameer, "Automated Categorization of Brain Tumor from MRI Using CNN features and SVM," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 8357–8369, 2021, doi: 10.1007/s12652-020-02568-w.

[24] R. Singh, A. Goel, and D. K. M. R. Raghuvanshi, "Brain tumor classification employing ICA and kernel-based support vector machine," *Signal, Image and Video Processing*, vol. 15, pp. 501–510, 2021, doi: 10.1007/s11760-020-01770-9.

[25] M. M. Badža and M. Č. Barjaktarović, "Classification of Brain Tumors from MRI Images Using a Convolutional Neural Network," *Applied Sciences*, vol. 10, no. 6, p. 1999, 2020, doi: 10.3390/app10061999.

[26] J. Y. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2242-2251, 2017, doi:10.1109/ICCV.2017.244.

[27] A. Rehman, S. Naz, M. I. Razzak, F. Akram, M. A. Inran, "A Deep Learning-Based Framework for Automatic Brain Tumors Classification Using Transfer Learning," *Circuits, Systems, and Signal Processing*, vol. 39, pp. 757–775, 2020, doi: 10.1007/s00034-019-01246-3.

[28] A. Seal, D. Bhattacharjee, and M. Nasipuri, "Predictive and probabilistic model for cancer detection using computer tomography images," *Multimedia Tools and Applications*, vol. 77, 2018, doi: 10.1007/s11042-017-4405-7.

[29] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8107-8116, 2020, doi: 10.1109/CVPR42600.2020.00813.

[30] M. Sajjad, S. Khan, K. Muhammad, W. Wu, A. Ullah, and S. Baik, "Multi-Grade Brain Tumor Classification using Deep CNN with Extensive Data Augmentation," *Journal of Computational Science*, vol. 30, pp. 174-182, 2018, doi: 10.1016/j.jocs.2018.12.003.

[31] A. Karnewar and O. Wang, "MSG-GAN: Multi-Scale Gradients for Generative Adversarial Networks," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7796-7805, 2020, doi: 10.1109/CVPR42600.2020.00782.

[32] A. Karnewar, O. Wang, and R. S. Iyengar, "MSG-GAN: Multi-Scale Gradient GAN for Stable Image Synthesis," *ArXiv preprint*, 2019, abs/1903.06048.

[33] J. Seetha and S. S. Raja, "Brain tumor classification using Convolutional Neural Networks," *Biomedical and Pharmacology Journal*, vol. 11, no. 3, pp. 1457-1461, 2018, doi: 10.13005/bpj/1511.

[34] M. Balasooriya and R. D. Nawarathna, “A sophisticated convolutional neural network model for brain tumor classification,” in *Proceedings of the 2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, Peradeniya, pp. 1-5, 2017, doi: 10.1109/ICIINFS.2017.8300364.

[35] P. Afshar, K. N. Plataniotis, A. Mohammadi, “Capsule networks for brain tumor classification based on MRI images and coarse tumor boundaries,” in *Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP)*, Athens, pp. 1368-1372, 2019, doi: 10.1109/ICIP.2018.8451379.

[36] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, “Generative Adversarial Networks,” *Advances in Neural Information Processing Systems*, vol. 27, 2014, doi: 10.1145/3422622.

[37] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation,” 2018, doi: arxiv:1710.10196v3.

[38] X. Huang and S. Belongie, “Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization,” *International Conference on Computer Vision*, pp. 1510-1519, 2017, doi: arXiv:1703.06868v2.

[39] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, “Improved Training of Wasserstein GANs,” pp. 5769-5779, 2017, doi: arXiv:1704.00028v3.

[40] D.P. Kingma, J. Ba, “Adam: A Method for Stochastic Optimization,” *International Conference on Learning Representations*, 2014.

Akshay Kumaar M

Received the B. Tech degree in Information Technology at Anna University - St. Joseph’s College of Engineering, Chennai, India in 2021. He is a Senior Machine Learning Engineer at BrainSightAI, a Neuroscience and Artificial Intelligence based company. He is highly passionate about AI in Healthcare and his research interests include signal processing & enhancement, self-supervised learning-based systems, and computer vision. He has contributed to multiple journal papers and has contributed to multiple research works that were presented at conferences like the Forum of European Neuroscience, and the Radiological Society of North America.

Duraimurugan Samiayya

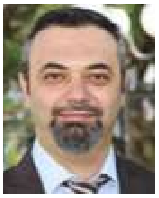
Obtained his Bachelor’s degree in Computer Science and Engineering from Noorul Islam College of Engineering, Tamilnadu, India and did his Master’s degree in Systems Engineering and Operations Research from College of Engineering Guindy, Anna University, Chennai, Tamilnadu India-25 and did his Ph.D Degree from Sathyabama Institute of science and technology, Chennai, Tamilnadu,india-119. Currently, he is an Associate Professor in the Department of Information Technology at St. Joseph’s College of Engineering, Chennai, Tamilnadu India-119. He has more than 15 years of teaching experience in Engineering College. His research interests include Congestion control in Multimedia Streaming and Image Processing.

P.M. Durai Raj Vincent

Received his B.E. and M.E. from Anna University, Chennai, India. He also received his Ph.D., from VIT University Vellore. He is presently working as an Associate Professor in the School of Information Technology and Engineering at Vellore Institute of Technology (VIT), India. He is having more than 15 years of teaching and research experience with over 90 Scopus indexed publications. His current research interest includes, Machine Learning and Data Analytics.

V. Rajinikanth

He is a Professor in Department of Computer Science, Division of Research and Innovation, Saveetha School of Engineering, SIMATS, Chennai 602105, Tamilnadu, India. He has published more than 150 papers and authored/edited 8 books in the field of medical data assessment. His main research interests includes; Heuristic algorithm based optimization, Image thresholding, Machine learning and Deep learning.



Seifedine Kadry

Professor Seifedine Kadry has a Bachelor 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. 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 ABET program evaluator for Engineering Tech. He is a Fellow of IET, Fellow of IETE, and Fellow of IACSIT. He is a distinguished speaker of IEEE Computer Society.

KoopaML: A Graphical Platform for Building Machine Learning Pipelines Adapted to Health Professionals

Francisco José García-Peñalvo[1], Andrea Vázquez-Ingelmo[1], Alicia García-Holgado[1], Jesús Sampedro-Gómez[2], Antonio Sánchez-Puente[2], Víctor Vicente-Palacios[3], P. Ignacio Dorado-Díaz[2], Pedro L. Sánchez[2] *

[1] GRIAL Research Group, University of Salamanca (Spain)
[2] Cardiology Department, Hospital Universitario de Salamanca, SACyL. IBSAL, Facultad de Medicina, University of Salamanca, and CIBERCV (ISCiii) (Spain)
[3] Philips Healthcare (Spain)

* Corresponding author. fgarcia@usal.es (F. J. García-Peñalvo), andreavazquez@usal.es (A. Vázquez-Ingelmo), aliciagh@usal.es (A. García-Holgado), jmsampedro@saludcastillayleon.es (J. Sampedro-Gómez), asanchezpu@saludcastillayleon.es (A. Sánchez-Puente), victor.vicente.palacios@philips.com (V. Vicente-Palacios), pidorado@saludcastillayleon.es (P. Ignacio Dorado-Díaz), plsanchez@saludcastillayleon.es (P. L. Sánchez).

Received 23 February 2022 | Accepted 1 June 2022 | Early Access 23 January 2023

Abstract

Machine Learning (ML) has extended its use in several domains to support complex analyses of data. The medical field, in which significant quantities of data are continuously generated, is one of the domains that can benefit from the application of ML pipelines to solve specific problems such as diagnosis, classification, disease detection, segmentation, assessment of organ functions, etc. However, while health professionals are experts in their domain, they can lack programming and theoretical skills regarding ML applications. Therefore, it is necessary to train health professionals in using these paradigms to get the most out of the application of ML algorithms to their data. In this work, we present a platform to assist non-expert users in defining ML pipelines in the health domain. The system's design focuses on providing an educational experience to understand how ML algorithms work and how to interpret their outcomes and on fostering a flexible architecture to allow the evolution of the available components, algorithms, and heuristics.

Keywords

Artificial Intelligence, HCI, Health Platform, Information System, Medical Data Management, Medical Imaging Management.

DOI: 10.9781/ijimai.2023.01.006

I. Introduction

Machine Learning (ML) has become a powerful approach to tackle complex tasks that involve analyzing significant amounts of data. Data-intensive contexts, such as the health domain, benefit directly from applying ML algorithms to their data, supporting tasks such as identifying patterns, clustering, classification, predictions, etc., that could become time- and resource-consuming if approached through manual paradigms. The application of ML to health data has proven its usefulness in specific challenges like diagnoses, disease detection, segmentation, assessment of organ functions, etc. [1] -[3].

However, applying ML approaches is not straightforward. More specifically, using them in sensitive domains (such as health) could be hazardous if practitioners do not fully understand the results derived from the models.

ML does not only consist of applying a set of pre-defined functions. It needs a deep understanding of the input data, the transformations that need to be performed to fit a model, the selection of a proper model, and its quality metrics before using trained models in production. Otherwise, the outputs could lead to wrong conclusions, losses, discrimination, and even negligence [4] -[7].

Therefore, it is necessary to balance data domain knowledge and ML expertise. While ML experts have a wealth of knowledge about ML algorithms, they can lack understanding regarding the input data. The same applies to health professionals; they have a profound knowledge of the data domain, but they would not obtain quality models without programming or ML skills.

In this scenario, it is necessary to provide practitioners with tools that alleviate this knowledge gap, enabling health professionals to implement ML pipelines and learn how, when, and why to apply specific models or functions to their data. This way, the introduction of ML in medical tasks could yield complementary support to automate and enhance decision-making processes without consuming an excessive quantity of resources and time.

In this context we pose the following research question:

RQ1. Which features can ease the application of ML algorithms in the medical context?

Please cite this article as: F. J. García-Peñalvo, A. Vázquez-Ingelmo, A. García-Holgado, J. Sampedro-Gómez, A. Sánchez-Puente, V. Vicente-Palacios, P. Ignacio Dorado-Díaz, P. L. Sánchez, "KoopaML: A Graphical Platform for Building Machine Learning Pipelines Adapted to Health Professionals", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 6, pp. 112-119, 2024, http://dx.doi.org/10.9781/ijimai.2023.01.006

Driven by this research question, we present a graphical platform (KoopaML) to offer intuitive and educational interfaces to build and run ML pipelines to tackle these challenges. The primary target audience of this platform is non-expert users interested in learning and applying ML models to their domain data. We followed a user-centered design approach to capture relevant requirements and necessities from potential user profiles involved in this context.

In addition, we focused on providing a flexible architecture to allow expert users to extend the platform's functionality through new custom algorithms, components, or new heuristics to guide the definition of ML pipelines.

In this paper, we describe the design process, the platform's architecture, its underlying processes, and the feedback obtained from experts regarding the first development stages of the system.

The rest of the work is structured as follows. Section 2 provides an overview of similar tools for learning and building ML and data science pipelines. Section 3 describes the methodology followed for eliciting requirements and the technologies employed to implement the platform. Section 4 details the platform's architecture and modular decomposition, while section 5 describes the implemented functionalities. Finally, sections 7 and 8 discuss the results and conclude the work, respectively.

II. Related Work

Plenty of helper tools has been developed due to the increasing popularity of ML. Specifically, there are three main categories: programming frameworks and libraries, platforms for experts and non-experts, and platforms that support learning and understanding regarding how ML algorithms and pipelines work.

The first category encloses several well-known programming libraries: TensorFlow [8], Apache Mahout [9], and other Python frameworks like PyTorch (https://pytorch.org/), Scikit-learn (https://scikit-learn.org/), or Keras.io (https://keras.io/). These libraries provide an abstraction layer to implement ML models, but they require programming skills to employ them properly.

The second category focuses on visual environments that assist users through intuitive interfaces in creating and defining ML pipelines. Weka, for instance, provides a collection of algorithms for data mining tasks. One of its environments enables users to define data streams by connecting nodes representing data sources, preprocessing tasks, evaluation methodologies, visualizations, or algorithms, among other [10], [11].

On the other hand, Orange Data Mining Field [12] allows the definition of data mining workflows, with several methodologies, operations, and visualizations available through a user-friendly interface. The possibility of introducing customized dashboards [12] -[14] to present the outcomes of ML tasks an extremely valuable feature to ease the comprehension of the pipeline stages. Another tool with similar features is Rapid Miner [15], which follows a node-and-link philosophy to specify and define ML workflows. These applications provide robust and complex features through intuitive interfaces and interaction methods, adding abstraction to programming libraries.

The last category refers to platforms whose primary goal is to offer a didactic experience and learning resources to ease understanding ML algorithms and workflows. Tools within this category provide user-friendly and simple graphical interfaces avoiding technical details. Examples include Machine Learning for Kids (https://machinelearningforkids.co.uk/) or LearningML (https://web.learningml.org/).

Although several solutions are developed to assist non-expert users in the definition of ML pipelines, it is difficult to adapt them to specific contexts with particular necessities and requirements. For these reasons, we opted to develop a customized tool focused on the provenance of an educative experience for health professionals that want to start applying ML models to support their tasks.

III. Methodology

A. Requirements Elicitation

We identified the main features and specifications of the platform through a requirements elicitation process. Specifically, we interviewed potential users and domain experts, including physicians, computer scientists, and managers.

The output of this process was the description of the platform's basic features:

1. Definition of ML pipelines
2. Execution of ML pipelines
3. Interpretation of ML results
4. Visualization of ML results
5. Data validation
6. Heuristics management

The first two features are related to implementing ML pipelines by connecting different tasks, including data preprocessing and cleaning, ML algorithms, and evaluation functions. The platform allows users to choose different ML algorithms and configure their parameters. Users can personalize their pipelines by connecting nodes and analyzing each step's intermediate results.

Features 3 and 4 are related to the outcomes obtained during the execution of the pipelines. Each stage will output new results, and these results need to be understood to gain insights. For these reasons, the platform needs to provide methods to convey and assist the interpretation of the pipeline outputs through explanations, annotations, and data visualizations. This process is vital because a wrong interpretation of the results could lead to useless results and lose all its potential benefits.

On the other hand, the quality of the training process not only depends on the algorithm's configuration but also on the quality of input data. The platform needs to support validation processes and emphasize cleaning and preprocessing functions before training ML models. This feature focuses on providing information regarding the applicability of the available algorithms to the input data and potential issues (missing values, data imbalance, data samples, data types, etc.).

The last functionality refers to applying heuristics to assist non-expert users in the definition of ML pipelines. The management of heuristics and recommendation rules should be flexible to support the evolution of the suitability of algorithms depending on the context. Therefore, the platform will allow the modification and addition of new heuristics to provide more flexibility and build customized rule-based recommenders.

During the elicitation process, two user roles were identified. This categorization of users is essential to adapt the functionalities depending on the role, as well as their privileges:

- Non-expert users. The primary users of the platform. Non-expert users (mainly physicians) who know the data domain are interested in IA and ML but don't have enough skills to create ML pipelines programmatically.
- AI experts. Experts will have access to the ML pipelines workspace, but they will also have privileges to define and modify heuristics to configure the recommendations or preferred workflows of the platform.

B. Development

As introduced, one of the main goals of this work is to provide a flexible platform with the capability to evolve to include discoveries in ML. Therefore, it is crucial to rely on flexible technologies and paradigms that support the reusability of components.

ML pipelines share common features and can be represented through abstract elements to leverage their commonalities and foster the reusability of core assets. We followed the software product line (SPL) paradigm and domain engineering to capture ML pipelines and tasks' commonalities and variability points and arrange the software components accordingly [16] - [20].

With this approach, it is possible to reuse these "building blocks" and modify/add new ones without impacting the rest of them. On the other hand, building each pipeline task as an independent component with well-defined inputs and outputs also meets the requirement of inspecting intermediate results and even executing pipelines step by step.

We materialized the variability of pipelines through SciLuigi (https://github.com/pharmbio/sciluigi), a wrapper for Spotify's Luigi Python library (https://github.com/spotify/luigi), which supports the definition of dynamic workflows avoiding hard-coded dependencies [21], [22].

C. Validation

To validate the first version of KoopaML, we carried out an expert judgement [23] validation with experts from the medical and AI fields. Three experts were recruited to thoroughly test the platform and seek for issues regarding its contents and interaction mechanisms.

The three participants are AI developers in the medical domain, so they were able to test the platform from the two perspectives.

IV. Architecture

Providing flexible and extensible architecture is crucial in this field, as approaches constantly improve and evolve. This section outlines the platform's architecture and the mechanisms employed to support the evolution of its components.

A. Modules

The architecture of KoopaML is based on different modules connected by information flows. One of the primary purposes of this design is to provide flexible pipelines with reusable components.

In this regard, we followed a domain engineering approach through the previously described requirements elicitation process with potential users and literature reviews.

Following this approach, we propose four general functional blocks that will interact and collaborate among them to provide support for the implementation of flexible ML workflows:

- User management module
- Heuristics management module
- Pipelines management module
- Tasks management module

The user management module provides the services related to authentication, sessions, and roles. The heuristics management module allows IA experts to modify the heuristics through a graphic interface. The pipelines management module provides a workspace to create ML pipelines using visual elements. Finally, the tasks management module defines the operations related to each ML pipeline potential stage.

Fig. 1 shows the schematic overview of the platform's architecture with the C4 model notation [24].

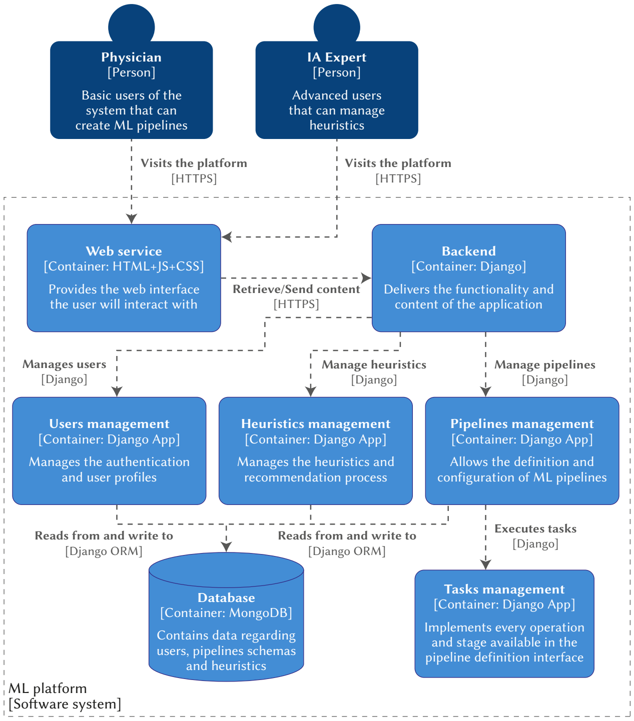


Fig. 1. Outline of the platform's architecture.

B. Pipelines' Structure

While the previous functional blocks provide flexibility to evolve the system's features, they still need fine-grained flexibility regarding the implementation of ML pipelines.

Following the software product line architecture paradigm [16] -[20], we divided ML pipelines into fine-grained tasks with well-defined inputs and outputs. Through this approach, the tasks management module acts as a repository of loosely coupled ML-related tasks, in which algorithms and operations can be added and modified without impacting the features of the remaining modules/tasks.

As explained in the methodology section, this encapsulation of ML tasks is achieved through the SciLuigi library. Fig. 2 outlines the structure of the pipelines following this approach.

Tasks are categorized following their high-level functionality (tasks related to data upload, data preprocessing, ML algorithms, or evaluation metrics). Then, more specific tasks are implemented; for example, within the "ML algorithms" category, we can find particular algorithms such as Naïve Bayes, Random Forest, Linear Regression, etc.

Users can instantiate nodes from each category and connect them through their inputs and outputs. These inputs and outputs are also categorized to ensure that information flows are compatible among the instantiated nodes.

The connection restrictions between nodes are implemented in the interface to ensure that the SciLuigi pipeline is correctly instantiated. With this method, the construction of the final SciLuigi pipeline is straightforward.

The simplified code in Fig. 3 outlines the implementation of a SciLuigi workflow through the pipeline specification defined by the user in the graphical interface. The main challenge was related to the dynamic connection of inputs and outputs. SciLuigi requires knowing the specific inputs/outputs names beforehand to connect them through explicit attribute value assignment. Lines 14-17 (Fig. 3) show how this issue was solved using the *setattr* and *getattr* Python methods, allowing dynamic access to class attributes.

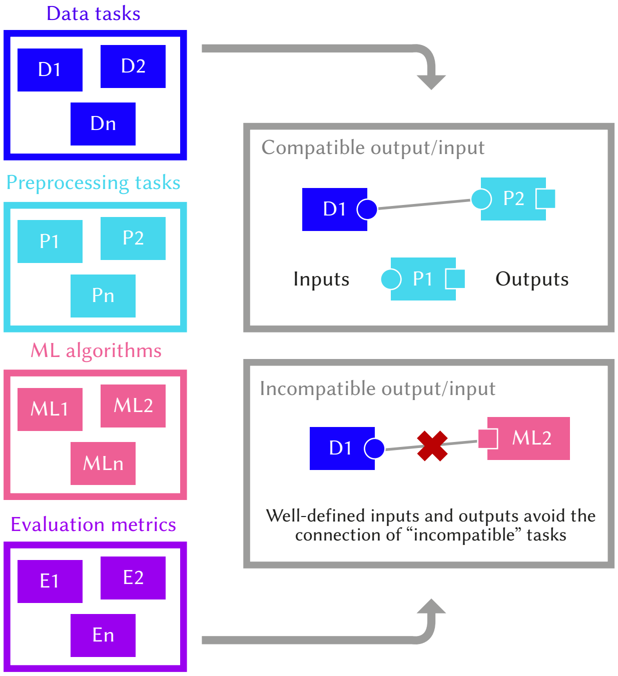
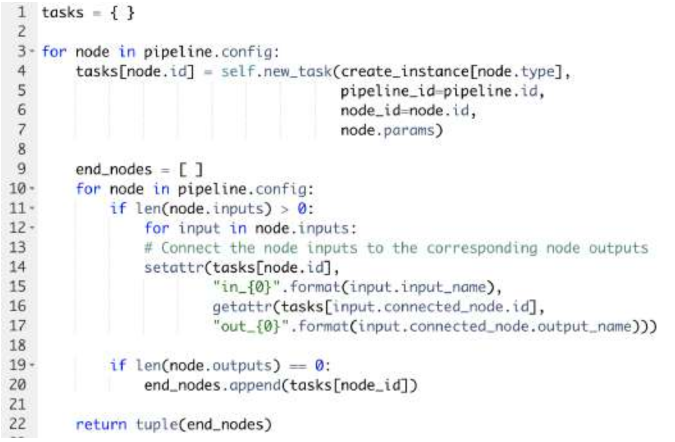


Fig. 2. Outline of programming approach. Each task contains its own logic and belongs to a specific category. Inputs and outputs compatibilities (in terms of information flows) are computed from each node's logic.



```
tasks = { }

for node in pipeline.config:
    tasks[node.id] = self.new_task(create_instance[node.type],
                                   pipeline_id=pipeline.id,
                                   node_id=node.id,
                                   node.params)

    end_nodes = [ ]
    for node in pipeline.config:
        if len(node.inputs) > 0:
            for input in node.inputs:
            # Connect the node inputs to the corresponding node outputs
            setattr(tasks[node.id],
                    "in_{0}".format(input.input_name),
                    getattr(tasks[input.connected_node.id],
                    "out_{0}".format(input.connected_node.output_name)))

        if len(node.outputs) == 0:
            end_nodes.append(tasks[node_id])

    return tuple(end_nodes)
```

Fig. 3. Algorithm to materialize a pipeline specification into a SciLuigi workflow (syntax simplified for readability).

This solution provides more flexibility and eases the addition and modification of new tasks, as the whole pipeline can be instantiated without hard-coding specific dependencies or class types.

V. KoopaML

A. Prototype

A prototype was developed and evaluated to complement the requirement elicitation process through a focus group. This methodology enabled us to capture more requirements and validate the platform's conceptual design before its implementation.

Fig. 4 shows a screenshot of the interface for defining ML pipelines through node-link structures.

The focus group involved different user profiles, including physicians and AI experts related to the health domain. The outcomes of this study can be consulted in [25]. The feedback was positive and helpful for starting the implementation of the tool.

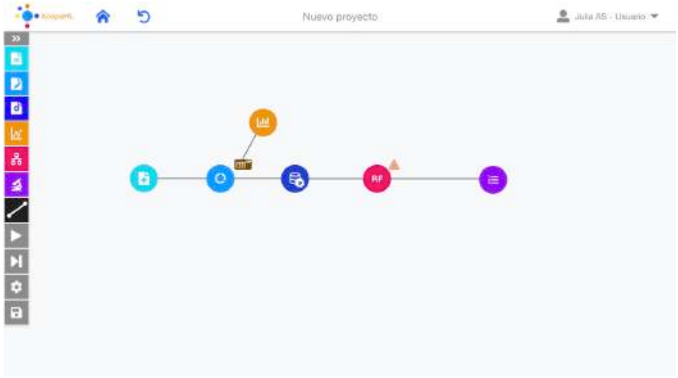


Fig. 4. Prototype of the workspace to define ML pipelines.

B. Functional System

As explained throughout this work, a crucial characteristic of the interface is that it should be simple to avoid overwhelming users with several complex concepts at once and robust to enable the definition of ML pipelines with enough detail. This section provides an overview of the interface proposal and the different features of the first version of KoopaML.

1. ML Pipelines

When creating a new project or pipeline, the system displays an empty workspace with a toolbar containing the tasks included in the ML workflow. As introduced in previous sections, tasks are divided into high-level categories to ease users' search of specific nodes (Fig. 5).

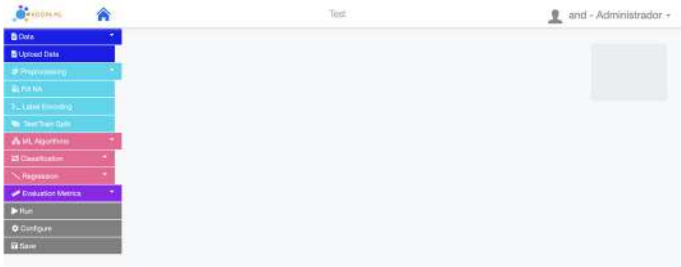
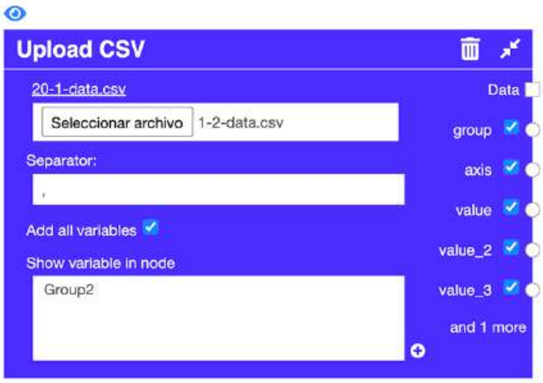


Fig. 5. New project workspace and available nodes.

Users can click on specific tasks or drag and drop them into the workspace to start configuring the pipeline. Fig. 6 shows the "Upload CSV" node. This node is particularly complex because several circumstances need to be considered when uploading data:

1. CSV files can be separated by different characters, such as commas or semicolons. For this reason, the node allows the configuration of varying separator values through a text input.
2. Some nodes could take as an input a single column (or a subset of them). That is why each dataset's columns need to be considered single outputs and be accessible to create connections among nodes.
3. The whole dataset is also considered a single output ("Data" socket in Fig. 6) to avoid multiple column connections and ease the data flows. This output includes the whole dataset (the set of all columns contained in the uploaded file).
4. Related to the previous point, some columns might be discarded from the dataset (i.e., columns that hold several missing values or aren't relevant to the problem). A checkbox beside each column allows users to select the columns that will be part of the dataset.
5. Finally, data related to the health domain can hold a significant quantity of variables. However, showing all variables as outputs in the "Upload CSV" node at once could impact the user experience. For this reason, a threshold has been configured to show only

the first five columns of a dataset, allowing the user to add the remaining columns through a multiple selection input. This way, users can focus only on variables that need explicit connections through their ML pipeline.


Upload CSV

20-1-data.csv

Seleccionar archivo 1-2-data.csv

Separator: ,

Add all variables ☑

Show variable in node: Group2

Data □

group ☑

axis ☑

value ☑

value_2 ☑

value_3 ☑

and 1 more

Fig. 6. A node for uploading CSV files.

Fig. 7 shows a simple ML workflow in which:

1. Categorical data is encoded through a Label Encoder.
2. The output from the label encoding process is then split into training and test sets. This node needs to know the variable to predict to perform the division of data. In this specific case, the variable "group" will be the one to be expected through this pipeline.
3. Training datasets are connected to a Random Forest classifier.
4. Finally, the trained model and the test datasets are connected to an evaluation node to measure the model's accuracy.

Users can execute the pipeline whenever they want by clicking the "Run" button, triggering the backend to build the pipeline by connecting tasks using the algorithm presented in Fig. 3. Once the pipeline has been executed, the workspace displays the results (or any error) individually in each node (bottom image of Fig. 7). Storing intermediate results leverages one of the main benefits of using SciLuigi, which is the possibility of re-running failed tasks individually without triggering the whole pipeline again.

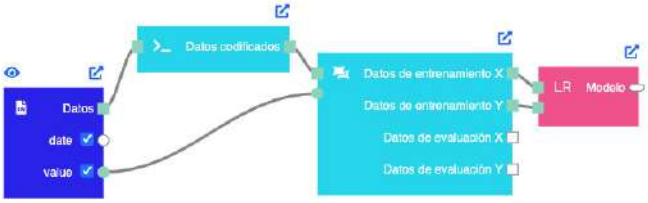


Fig. 7. Execution and results of an ML pipeline. Intermediate results of each node can be consulted by clicking on the top-right icon of each node.

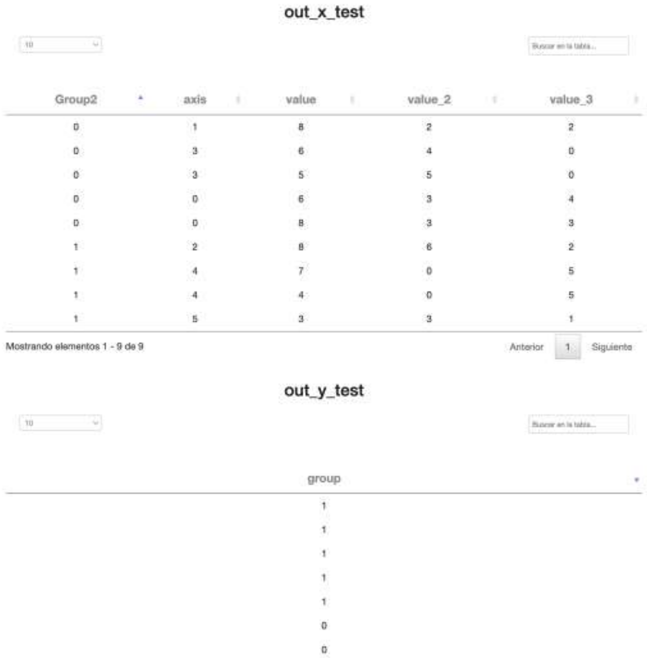
Through this approach, intermediate results can also be inspected individually. On the one hand, Fig. 8 displays the intermediate results from the test/train splitting node. This node yields four results: test and training datasets separated by the column to predict. The fig. 8 shows two of these intermediate results (the test datasets).

Evaluation metrics are also treated as intermediate results. In this case, the measurement of the accuracy of the trained model yielded 33% of correct predictions (Fig. 9).

2. Data Validation

Data validation and exploratory analysis are crucial steps when building successful ML pipelines. If data is not properly inspected and preprocessed, trained models could yield useless results. KoopaML provides a summary screen to assist users in the exploration process. This section is divided into three main blocks.

The first block provides a table view of the whole dataset. This view allows users to see all columns and rows of the uploaded data files and navigate through them in detail (Fig. 10).


out_x_test

10

Buscar en la tabla...

Group2	axis	value	value_2	value_3
0	1	8	2	2
0	3	6	4	0
0	3	5	5	0
0	0	6	3	4
0	0	8	3	3
1	2	8	6	2
1	4	7	0	5
1	4	4	0	5
1	5	3	3	1

Mostrando elementos 1 - 9 de 9

Anterior 1 Siguiente

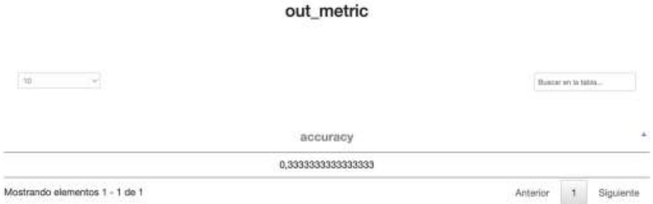
out_y_test

10

Buscar en la tabla...

group
1
1
1
1
1
0
0

Fig. 8. Results were derived from splitting the uploaded data into test and training datasets.


out_metric

10

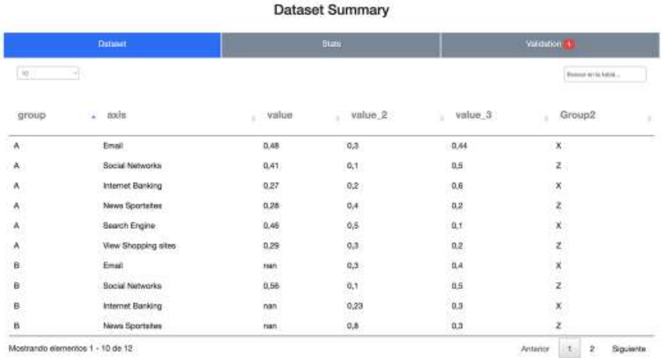
Buscar en la tabla...

accuracy
0,3333333333333333

Mostrando elementos 1 - 1 de 1

Anterior 1 Siguiente

Fig. 9. Accuracy of the trained model. Note that low accuracy is related to the small dataset that illustrates the system's functionalities.


Dataset Summary

Dataset | Stats | Validation 1

10

Buscar en la tabla...

group	axis	value	value_2	value_3	Group2
A	Email	0,48	0,3	0,44	X
A	Social Networks	0,41	0,1	0,5	Z
A	Internet Banking	0,27	0,2	0,6	X
A	News Sportsites	0,28	0,4	0,2	Z
A	Search Engine	0,46	0,5	0,1	X
A	View Shopping sites	0,29	0,3	0,2	Z
B	Email	nan	0,3	0,4	X
B	Social Networks	0,56	0,1	0,5	Z
B	Internet Banking	nan	0,23	0,3	X
B	News Sportsites	nan	0,8	0,3	Z

Mostrando elementos 1 - 10 de 12

Anterior 1 2 Siguiente

Fig. 10. Results were derived from splitting the uploaded data into test and training datasets.

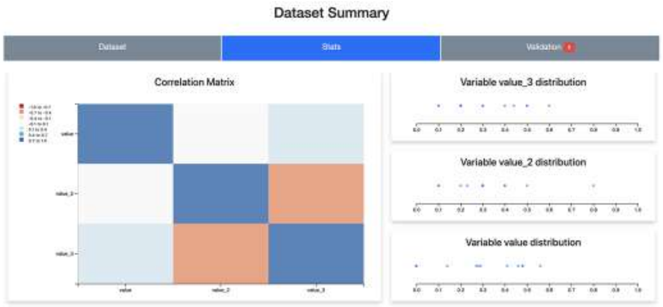
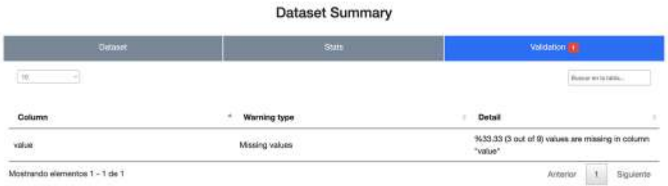

Dataset Summary
Dataset
Stats
Validation
Correlation Matrix
value
value_2
value_3
Variable value_3 distribution
Variable value_2 distribution
Variable value distribution

Fig. 11. Information dashboard of the input dataset characteristics.

The second summary block is a data dashboard in which practical data details, such as the distribution of values, data types, number of missing values, or a correlation matrix, are presented visually to ease the analysis of the dataset characteristics (Fig. 11). The dashboard is automatically generated and tailored according to the user needs [26] - [28].

Finally, the last block is focused on alerting users regarding potential issues of the dataset (Fig. 12), such as columns with significant quantities missing values, mixed data types, unbalanced categories, etc. Users are encouraged to consider or solve these issues through this feature before using the dataset in a pipeline.



Column	Warning type	Detail
value	Missing values	%33.33 (3 out of 9) values are missing in column "value"

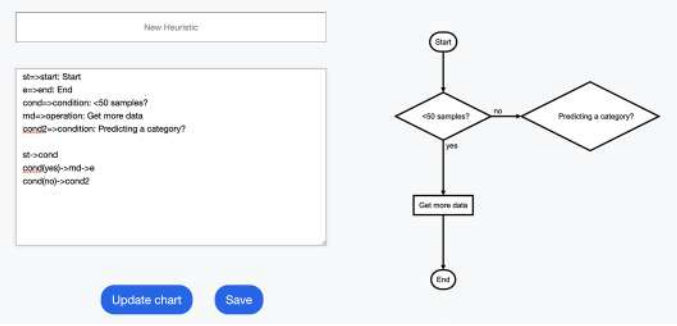
Fig. 12. Validation screen.

3. Heuristics Management

As explained before, one of the goals of the platform users is to learn from the experience of developing pipelines and build skills related to the application of ML. However, this learning experience needs to be guided by expert knowledge.

We have tackled this challenge through the definition and management of custom heuristics. KoopaML allows expert users to design heuristics in graphical decision trees to yield recommendations and guide the implementation of pipelines.

Heuristics are represented through the DSL provided by the flowchart.js (https://github.com/adrai/flowchart.js) library. This library allows textual and graphical representation of flow charts, providing a fine-grain manipulation of heuristics and rule-based recommendations (Fig. 13).



```
st=>start: Start
e=>end: End
cond=>condition: <50 samples?
md=>operation: Get more data
cond2=>condition: Predicting a category?

st->cond
cond(yes)->md->e
cond(no)->cond2
```

Fig. 13. Example of the definition of a heuristic.

VI. Expert Validation

The results of the expert validation were favorable. Overall, the platform was rated as useful to overcome the difficulties of creating ML pipelines in a medical context.

Regarding the issues encountered, apart from minor bugs that were fixed, the following can be highlighted:

- **Error reports.** The experts pointed out the possibility of having a variety of errors related to the execution of the pipeline. In the current version of KoopaML, these errors were displayed through tooltips associated to each node. However, experts indicated that it might be useful to have an unified report listing every error or warning raised during the execution of the pipeline.
- **Model metrics.** KoopaML allows the computation of different metrics to validate the trained models. For this matter, the user needs to select and connect every metric they want to calculate. This could be time-consuming if several metrics are to be analyzed. In this sense, the experts advised the possibility of unifying every metric on a node, and let the user select the metrics directly from there instead of carrying out the selection one by one.
- **Data visualizations.** The data summary presented in the previous section was highly valued by the experts. Following this idea, they recommended implementing a dashboard with visualizations related to model metrics as well.
- **Cross validation.** The experts pointed out that, in practice, they use cross validation [29], and thus, that the platform should support this approach.

Other comments were related to the addition of a wide set of algorithms and metrics, as well the possibility of configuring the hyperparameters of the algorithms through the interface.

VII. Discussion

This work presents the first version of KoopaML: a platform for automating and learning the definition of ML pipelines. We followed a user-centered approach for the design and development process, considering the primary goal of the system: to ease the application of ML for non-specialized users.

This version has been subject to iterative development with continuous feedback from experts. For instance, the "Upload CSV" node design shown in Fig. 6 resulted from different evaluations in which domain experts exposed issues encountered or potential improvements when uploading their domain data.

Although there are commercial tools that tackle the automation of these processes, the specific requirements that arise from the medical context asked for a customized platform that aligns with the necessities of end-users (in this case, physicians with lack of data science skills but that are interested in applying ML).

On the other hand, another related benefit of the customized tool is implementing communication mechanisms among other already developed devices for the cardiology department at the University Hospital of Salamanca [30]. Connecting different platforms would foster the creation of a technological ecosystem [31] with powerful and transparent data management and data science features adapted to the health sector requirements.

The platform's architecture is designed to allow flexibility and evolution due to the changing nature of AI and ML methods. The abstraction of pipelines into tasks with well-defined inputs and outputs has facilitated the user interface design and the final implementation of the workflows through libraries such as SciLuigi, matching the same node-link structures.

In addition to the workspace for instantiating pipelines, the platform also provides an interface to support the exploratory analysis of data. This interface was included after the evaluation of the platform by expert users, who asked for more feedback related to the input data.

Finally, one novel feature of KoopaML is the heuristics management module. This module enables the definition of heuristics through a DSL and its graphical representation. Heuristics can be stored to rely on different knowledge bases depending on the data domain, for example. The dynamic heuristic definition fosters the flexibility of the recommendations and guided support provided within the workspace during the implementation of ML pipelines. Moreover, their structured format allows the inclusion of external heuristics from other knowledge bases stores [32], [33].

Regarding the expert validation, the results were highly valuable and useful to set the foundations of new improvements and features, as well as to identify minor bugs. Having experts from both AI and medical fields enabled the identification of issues and shortcomings of the current version of the platform. For these reasons, we will continue performing this kind of evaluations, as they provide insights related to theoretical concepts that will be difficult to reach with lay users.

Following the research question posed in the introduction and the results of the expert validation, the platform has been developed taking into account the necessities of the medical domain. The implementation of an interface with simple and visual mechanisms (such as drag and drop or visually connecting two nodes to instantiate a pipeline) set the foundations for a platform that can be used by non-expert users.

On the other hand, the development of the heuristics management module will also allow the definition of recommendations that could be adapted to any kind of user. These features will provide additional assistance while creating and interpreting ML pipelines.

VIII. Conclusions

This work describes the design process, architecture, and features of KoopaML: a graphical platform for building machine learning pipelines adapted to health professionals.

The platform has been designed to support the evolution and addition of new tasks related to ML pipelines through abstraction mechanisms. The abstraction of tasks has allowed simplifying the user interface and the automatic implementation of the graphically instantiated pipelines.

KoopaML assists users in the definition of ML pipelines, execution of ML pipelines, interpretation and visualization of ML results, data validation, and heuristics management.

Future research lines will involve further expert validations of the platform, as well as in-depth user tests to measure the usability, ease of use, and effectiveness of the tool.

Acknowledgment

This research work has been supported by the Spanish *Ministry of Education and Vocational Training* under an FPU fellowship (FPU17/03276). This research was partially funded by the Spanish Government Ministry of Economy and Competitiveness through the DEFINES project grant number (TIN2016-80172-R) and the Ministry of Science and Innovation through the AVisSA project grant number (PID2020-118345RB-I00). This work was also supported by national (PI14/00695, PIE14/00066, PI17/00145, DTS19/00098, PI19/00658, PI19/00656 Institute of Health Carlos III, Spanish Ministry of Economy and Competitiveness and co-funded by ERDF/ESF, "Investing in your future") and community (GRS 2033/A/19, GRS 2030/A/19, GRS 2031/A/19, GRS 2032/A/19, SACYL, Junta Castilla y León) competitive grants.

References

[1] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," (in eng), *Med Image Anal,* vol. 42, pp. 60-88, Dec 2017, doi: 10.1016/j.media.2017.07.005.

[2] S. González Izard, R. Sánchez Torres, Ó. Alonso Plaza, J. A. Juanes Méndez, and F. J. García-Peñalvo, "Nextmed: Automatic Imaging Segmentation, 3D Reconstruction, and 3D Model Visualization Platform Using Augmented and Virtual Reality," (in eng), *Sensors (Basel),* vol. 20, no. 10, p. 2962, 2020, doi: 10.3390/s20102962.

[3] S. G. Izard, J. A. Juanes, F. J. García Peñalvo, J. M. G. Estella, M. J. S. Ledesma, and P. Ruisoto, "Virtual Reality as an Educational and Training Tool for Medicine," *Journal of Medical Systems,* vol. 42, no. 3, p. 50, 2018/02/01 2018, doi: 10.1007/s10916-018-0900-2.

[4] J. C. Weyerer and P. F. Langer, "Garbage in, garbage out: The vicious cycle of ai-based discrimination in the public sector," in *Proceedings of the 20th Annual International Conference on Digital Government Research*, Dubai, United Arab Emirates, 2019, pp. 509-511, doi: https://doi.org/10.1145/3325112.3328220.

[5] X. Ferrer, T. van Nuenen, J. M. Such, M. Coté, and N. Criado, "Bias and Discrimination in AI: a cross-disciplinary perspective," *IEEE Technology and Society Magazine,* vol. 40, no. 2, pp. 72-80, 2021, doi: 10.1109/MTS.2021.3056293.

[6] S. Hoffman, "The Emerging Hazard of AI-Related Health Care Discrimination," *Hastings Center Report,* vol. 51, no. 1, pp. 8-9, 2021, doi: 10.1002/hast.1203.

[7] S. Wachter, B. Mittelstadt, and C. Russell, "Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI," *Computer Law & Security Review,* vol. 41, p. 105567, 2021, doi: 10.2139/ssrn.3547922.

[8] M. Abadi *et al.*, "TensorFlow: A System for Large-Scale Machine Learning," in *12th USENIX Symposium on Operating Systems Design and Implementation OSDI 16.* Savannah, GA: USENIX Association, 2016, pp. 265-283.

[9] R. Anil *et al.*, "Apache Mahout: Machine Learning on Distributed Dataflow Systems," *Journal of Machine Learning Research,* vol. 21, no. 127, pp. 1-6, 2020. [Online]. Available: https://jmlr.org/papers/v21/18-800.html.

[10] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, and I. H. Witten, "Weka-A Machine Learning Workbench for Data Mining," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach Eds. Boston, MA: Springer, 2009.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explor. Newsl.,* vol. 11, no. 1, pp. 10–18, 2009, doi: 10.1145/1656274.1656278.

[12] A. Vázquez-Ingelmo, F. J. García-Peñalvo, and R. Therón, "Information Dashboards and Tailoring Capabilities - A Systematic Literature Review," *IEEE Access,* vol. 7, pp. 109673-109688, 2019, doi: 10.1109/ACCESS.2019.2933472.

[13] A. Sarikaya, M. Correll, L. Bartram, M. Tory, and D. Fisher, "What Do We Talk About When We Talk About Dashboards?," *IEEE Transactions on Visualization Computer Graphics,* vol. 25, no. 1, pp. 682 - 692, 2018, doi: 10.1109/TVCG.2018.2864903.

[14] S. Few, *Information dashboard design.* Sebastopol, CA, USA: O'Reilly Media, 2006.

[15] S. Land and S. Fischer, "Rapid miner 5," *Rapid-I GmbH,* 2012.

[16] J. Bosch, "From software product lines to software ecosystems," in *SPLC*, 2009, vol. 9, pp. 111-119.

[17] L. Chen, M. Ali Babar, and N. Ali, "Variability management in software product lines: a systematic review," 2009.

[18] P. Clements and L. Northrop, *Software product lines.* Addison-Wesley Boston, 2002.

[19] C. Kästner, S. Apel, and M. Kuhlemann, "Granularity in software product lines," in *2008 ACM/IEEE 30th International Conference on Software Engineering*, 2008: IEEE, pp. 311-320.

[20] J. Van Gurp, J. Bosch, and M. Svahnberg, "On the notion of variability in software product lines," in *Proceedings Working IEEE/IFIP Conference on Software Architecture*, 2001: IEEE, pp. 45-54.

[21] S. Lampa, J. Alvarsson, and O. Spjuth, "Towards agile large-scale predictive modelling in drug discovery with flow-based programming design principles," *Journal of Cheminformatics,* vol. 8, no. 1, p. 67, 2016/11/24 2016, doi: 10.1186/s13321-016-0179-6.

[22] S. Lampa, M. Dahlö, J. Alvarsson, and O. Spjuth, "SciPipe: A workflow library for agile development of complex and dynamic bioinformatics pipelines," *GigaScience,* vol. 8, no. 5, 2019, doi: 10.1093/gigascience/giz044.

[23] R. Skjong and B. H. Wentworth, "Expert judgment and risk perception," in *Proceedings of the Eleventh (2001) International Offshore and Polar Engineering Conference (Stavanger, Norway, June 17-22, 2001)*: International Society of Offshore and Polar Engineers, 2001.

[24] S. Brown. "The C4 Model for Software Architecture." https://c4model.com/ (accessed 16-05-2022).

[25] A. García-Holgado *et al.*, "User-Centered Design Approach for a Machine Learning Platform for Medical Purpose," Cham, 2021: Springer International Publishing, in Human-Computer Interaction, pp. 237-249.

[26] A. Vázquez Ingelmo, A. García-Holgado, F. J. García-Peñalvo, and R. Therón Sánchez, "A Meta-modeling Approach to Take into Account Data Domain Characteristics and Relationships in Information Visualizations," in *9th World Conference on Information Systems and Technologies*, Azores, Portugal, Á. Rocha, H. Adeli, G. Dzemyda, F. Moreira, and A. M. R. Correia, Eds., 2021, vol. 2: Springer Nature, in Trends and Innovations in Information Systems and Technologies, WorldCIST 2021, pp. 570-580, doi: 10.1007/978-3-030-72651-5_54. [Online]. Available: http://hdl.handle.net/10366/145626

[27] A. Vázquez-Ingelmo, A. García-Holgado, F. J. García-Peñalvo, and R. Therón, "Proof-of-concept of an information visualization classification approach based on their fine-grained features," *Expert Systems,* vol. n/a, no. n/a, p. e12872, In Press, doi: https://doi.org/10.1111/exsy.12872.

[28] A. Vázquez-Ingelmo, F. J. García-Peñalvo, and R. Therón, "Taking advantage of the software product line paradigm to generate customized user interfaces for decision-making processes: a case study on university employability," *PeerJ Computer Science,* vol. 5, p. e203, 2019/07/01 2019, doi: 10.7717/peerj-cs.203.

[29] C. Schaffer, "Selecting a classification method by cross-validation," *Machine Learning,* vol. 13, no. 1, pp. 135-143, 1993.

[30] F. García-Peñalvo *et al.*, "Application of Artificial Intelligence Algorithms Within the Medical Context for Non-Specialized Users: the CARTIER-IA Platform," *International Journal of Interactive Multimedia and Artificial Intelligence,* vol. 6, no. 6, 2021.

[31] A. García-Holgado and F. J. García-Peñalvo, "Validation of the learning ecosystem metamodel using transformation rules," *Future Generation Computer Systems,* vol. 91, pp. 300-310, 2019/02/01/ 2019, doi: https://doi.org/10.1016/j.future.2018.09.011.

[32] A. Martínez-Rojas, A. Jiménez-Ramírez, and J. Enríquez, "Towards a Unified Model Representation of Machine Learning Knowledge," presented at the Proceedings of the 15th International Conference on Web Information Systems and Technologies, Vienna, Austria, 2019. [Online]. Available: https://doi.org/10.5220/0008559204700476.

[33] C. Kumar, M. Käppel, N. Schützenmeier, P. Eisenhuth, and S. Jablonski, "A Comparative Study for the Selection of Machine Learning Algorithms based on Descriptive Parameters," in *Proceedings of the 8th International Conference on Data Science, Technology and Applications. Volume 1. DATA*, Prague, Czech Republic, 2019, pp. 408-415, doi: 10.5220/0008117404080415.

Francisco José García-Peñalvo

He received the degrees in computing from the University of Salamanca and the University of Valladolid, and a Ph.D. from the University of Salamanca (USAL). He is Full Professor of the Computer Science Department at the University of Salamanca. In addition, he is a Distinguished Professor of the School of Humanities and Education of the Tecnológico de Monterrey, Mexico. Since 2006 he is the head of the GRIAL Research Group GRIAL. He is head of the Consolidated Research Unit of the Junta de Castilla y León (UIC 81). He was Vice-dean of Innovation and New Technologies of the Faculty of Sciences of the USAL between 2004 and 2007 and Vice-Chancellor of Technological Innovation of this University between 2007 and 2009. He is currently the Coordinator of the PhD Programme in Education in the Knowledge Society at USAL. He is a member of IEEE (Education Society and Computer Society) and ACM.

Andrea Vázquez-Ingelmo

Andrea Vázquez-Ingelmo received the bachelor's degree in computer science from the University of Salamanca, Salamanca, in 2016 and the master's degree in computer science from the same university in 2018. She is a member of the Research Group of Interaction and eLearning (GRIAL), where she is pursuing her PhD degree in computer science. Her area of research is related to human-computer interaction, software engineering, data visualization and machine learning applications.

Alicia García-Holgado

She received the degree in Computer Sciences (2011), a M.Sc. in Intelligent Systems (2013) and a Ph.D. (2018) from the University of Salamanca, Spain. She is member of the GRIAL Research Group of the University of Salamanca since 2009. Her main lines of research are related to the development of technological ecosystems for knowledge and learning processes management in heterogeneous contexts, and the gender gap in the technological field. She has participated in many national and international R&D projects. She is a member of IEEE (Women in Engineering, Education Society and Computer Society), ACM (and ACM-W) and AMIT (Spanish Association for Women in Science and Technology).

Jesús Sampedro-Gómez

Jesús Sampedro-Gómez is industrial engineer by the Universidad Politécnica of Madrid. He is also a PhD student by the University of Salamanca and works as data scientist in the Cardiology Department of the University Hospital of Salamanca.

Antonio Sánchez-Puente

Antonio Sánchez Puente, PhD, is a junior researcher from CIBER working at the cardiology department of the University Hospital of Salamanca as a data scientist. He was awarded his doctorate in physics by the University of Valencia for his study of gravity theories before turning his career around the application of artificial intelligence in medicine.

Víctor Vicente-Palacios

Víctor Vicente-Palacios holds a PhD from the University of Salamanca in Statistics and works as a Data Scientist at Philips Healthcare in the area of AI applied to Medicine. He is also an alumnus of the Data Science for Social Good program (University of Chicago) and organizer of PyData Salamanca.

P. Ignacio Dorado-Díaz

P. Ignacio Dorado-Díaz holds a PhD from the University of Salamanca in Statistics. He is currently working as a research coordinator in the Cardiology Department of the Hospital de Salamanca in addition to being a professor at the Universidad Pontificia de Salamanca.

Pedro Luis Sánchez

Pedro Luis Sánchez holds a doctorate in medicine from the University of Salamanca. He is currently the head of the Cardiology Department at the University Hospital of Salamanca, in addition to being a professor at the University of Salamanca.

GRASE: Granulometry Analysis With Semi Eager Classifier to Detect Malware

Mahendra Deore[1]*, Manoj Tarambale[2], Jambi Ratna Raja Kumar[3], Sachin Sakhare[4]

[1] Department of Computer Engineering, MKSSS's Cummins College of Engineering for Women, Pune-411052 (India)
[2] Electrical Engineering Department, PVG's COET and GKP IOM, Pune- 411009 (India)
[3] Department of Computer Engineering, Genba Sopanrao Moze College of Engineering, Pune-411045 (India)
[4] Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune 411048 (India)

* Corresponding author. mdeore83@gmail.com

Received 26 February 2023 | Accepted 13 November 2023 | Early Access 12 December 2023

Abstract

Technological advancement in communication leading to 5G, motivates everyone to get connected to the internet including 'Devices', a technology named Web of Things (WoT). The community benefits from this large-scale network which allows monitoring and controlling of physical devices. But many times, it costs the security as MALicious softWARE (MalWare) developers try to invade the network, as for them, these devices are like a 'backdoor' providing them easy 'entry'. To stop invaders from entering the network, identifying malware and its variants is of great significance for cyberspace. Traditional methods of malware detection like static and dynamic ones, detect the malware but lack against new techniques used by malware developers like obfuscation, polymorphism and encryption. A machine learning approach to detect malware, where the classifier is trained with handcrafted features, is not potent against these techniques and asks for efforts to put in for the feature engineering. The paper proposes a malware classification using a visualization methodology wherein the disassembled malware code is transformed into grey images. It presents the efficacy of Granulometry texture analysis technique for improving malware classification. Furthermore, a Semi Eager (SemiE) classifier, which is a combination of eager learning and lazy learning technique, is used to get robust classification of malware families. The outcome of the experiment is promising since the proposed technique requires less training time to learn the semantics of higher-level malicious behaviours. Identifying the malware (testing phase) is also done faster. A benchmark database like malimg and Microsoft Malware Classification challenge (BIG-2015) has been utilized to analyse the performance of the system. An overall average classification accuracy of 99.03 and 99.11% is achieved, respectively.

Keywords

Granulometry Analysis, Malware, Semi Eager Classification (Semi-E), Static and Dynamic Analysis.

DOI: 10.9781/ijimai.2023.12.002

I. Introduction

Malicious software is baleful to all the devices connected to an internet, irrespective of the platform i.e., windows on laptop or android in a mobile. Presently android applications are growing exponentially to the scale of approximately 5 million apps in Google play as of May 2021 surpassing 2.99 million in the year 2020[1]. In parallel, malicious apps are also increasingly creating threats to mobile based financial transactions, taking control over mobile cameras, and misusing the same. According to the survey done by AV-TEST institute, there are approximately 1214.76 million malicious apps in the year 2021. Everyday AV-TEST registers approximately 350,000 new malicious apps and potentially unwanted applications[2]. To cope with the security threats various techniques for malware detection have been proposed by the researchers. It has been found that Machine Learning (ML) based detection technique is one of the efficient methods to opt for Malware Detection System (MDS). ML based MDS is comprehensive, detects malware accurately and less dependency on human experts which is normally required in traditional MD techniques. Thus, ML techniques are found to be more suitable for present scenarios where malicious software is increasing day by day.

Traditionally, the ML technique is feature vector based in which important characteristics of malware are extricated and used for identifying the same in a real time system. Static and dynamic are the two primary feature sources which describe malware characteristics.

[1] https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/

[2] https://www.av-test.org/en/statistics/malware/

Please cite this article as:
M. Deore, M. Tarambale, J. R. R. Kumar, S. Sakhare, "GRASE: Granulometry Analysis With Semi Eager Classifier to Detect Malware", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 6, pp. 120-134, 2024, http://dx.doi.org/10.9781/ijimai.2023.12.002

Both the analysis techniques can be applied to various kinds of executable files like PE, ELF, DEX etc., of different processors and operating systems (OS) such as Microsoft Windows, Linux, Android, etc.

The Static Analysis (SA) of malware software is done without the code being executed [1]. In SA, features are extracted after unpacking the executable in advance. Examples of static features are, OPCODE (Operation CODE) frequency distribution, control or data flow graph, syntactic library call, byte-sequence n-grams, string signature etc. Static features can be extracted from an image, sound, native code, and byte code.

In Dynamic Analysis (DA) run time behavior of malware executable is monitored. Due to this code is executed by using controlled environments like sandbox, emulator, simulator, virtual machine, etc. Dynamic features are extracted from sensitive function calls, variable value tracking, code execution path, log records and other behavior of the code when the same is being executed. Normally, analysis of the code is done by different tools like Process monitor (.pmon), Capture Bat, poison IVY, etc. The report generated from the tool is extensive and in depth, requiring human interpretation. Automated analysis of the report can be achieved but with huge computational complexities. Therefore, it is time consuming [2].

SA is preferred due to reliable detection efficiency, full code coverage, unperceived by malware code and simplicity to generate generic fingerprint of the malware code. In MDS one can extract SA features which will be input to machine learning algorithms for training the system. But due to the introduction of Deep Neural Network (DNN) architecture, huge amounts of data, maybe a vector matrix or an image, can be given as an input for training the network. The next section discusses MDS approach based on an image.

A. Visual Analytic Technique to Solve Challenges in MDS

The malware executable is a binary file, and it can be represented as strings of ones and zeros. A string being an array of hexadecimal values can be reshaped in matrix form and can be viewed as a grayscale image. The technique is Visual Analytic Techniques (VAT) and thus, MD can be put into an image recognition problem. VAT is mostly used for documents where the files are huge and for image analysis where the data is massive. Therefore, the technique is suitable to be used for computer security, as malware attacks are almost always in thousands at any given time.

There are four motivational factors to select VAT. Firstly, the image classification techniques are mature and furthermore it is faster [3]. The second point deals with the mindset of a malware developer. They work hard to hide the code and simultaneously come up with variants of malware; normally they just utilize the old code. Meghna Dhalaria et al. [4] use a robust set of features from static and dynamic malware analysis for creating two datasets i.e., binary, and multiclass (family) classification datasets. In such conditions, for a single malware family, the deviation between two or more gray scale images will be very less which opens a huge number of algorithms based on Similarity Mining Machine Learning (SMML). Thirdly, neither disassembly nor code execution is required for classification based on visualization. Finally, VAT for MD does not require code analysis and it is resilient to obfuscation techniques like polymorphism, packing and section encryption. Samples of malware images are given in Fig. 4. From the images we can conclude that the texture of images of malware families is different. So, texture-based analysis will be a more suitable solution for MD. The next section describes a texture-based approach which is explored in this paper.

B. Granulometry Based Texture Analysis of a Gray Scale Malware Image

Granularity is the random optical texture of an image. In an image, pixels are considered as *'Grains'*. Extracting spatial features such as shape and size from the image is more complex as compared to extracting textural features which does not require any type of segmentation. Researchers have proved that classification accuracy significantly increases if only textural information of an image is taken into consideration. Texture analysis methods are grey level co-occurrence matrix (GLCM), Markov random fields, Laplace filters, discrete wavelet transformation, fractal analysis and GRanulometric Analysis (GRA).

GRA is lesser known, but its significance was proved by Kupidura [5] and Skullmowska [6]. Morphological closing and opening operations as well as measuring the difference between successive images is the base of GRA. This characteristic permits the quantification of different size particles in an image [7]. Haas et al. [8] introduced this technique. Dougherty et al. [9] introduced methods of local analysis which allows assignment of texture values to individual pixels. This feature motivated researchers to use this technique for satellite image analysis. On the similar line we used GRA because in the gray scale malware image, each and every pixel is a malware code byte which is important for the malware family analysis. Thus, GRA provides pixel level analysis.

To the best of our knowledge GRA has not been used till date for analyzing malware images therefore motivating us to work on the same. In ML, feature extraction block and classifier block, both are of utmost importance. Therefore, after finalizing the textural feature for the proposed work the next part describes the classifier used for MDS.

C. Semi Eager (SemiE) Classifier

The task of the classifier is to accurately predict a malware family group of the malware input captured by the system. The learning process is the base of the classifier. Lazy and Eager learning are two types of techniques used in machine learning.

Conditional Random Field (CRF), Support Vector Machine (SVM) and Artificial Neural Network (ANN) are Eager Learning (EL) algorithms. The disadvantages of EL are as follows. The first is the high training time cost e.g., training time for SVM is $O(n^3)$, where n is the number of training instances. The second is the drifting and information loss, leading to over fitting or under fitting risk. The reason is that it computes global models after analyzing prediction query. Finally, there is impact of global distribution on full dataset instead of local behavior of unpredicted targets.

Lazy Learning (LL) or delayed learning is instance based where it memorizes present training examples and waits for the new instance to occur. Thus, in this method instead of estimating the entire instance space it estimates only the different and local instances. Locally Weighted Regression (LWR) and KNN calculate distance to each training example for predicting new instances, thus it follows LL approach. A. Zakai et al. [10] put forth that for the convergence of ML models, local behavior is important. LL can commit a plentiful set of hypotheses. The drawback of LL is, despite no training overhead, prediction time complexity is more i.e., $O(n)$, where n is the number of training examples. The SemiE learning algorithm overcomes the disadvantages of both the training techniques without any compromise of the advantages.

This paper proposes a GRASP MDS architecture consisting of a SemiE learning network for accurate detection and classification of malware families making use of image-based approaches. A benchmark database from Kaggle and malimg is utilized to assess the performance of the system. Features will be extracted from the greyscale image of different malware families and will be trained using SemiE learning.

The primary contribution of the research work is as follows:

1. To provide critical overview of related work based on VAT (image-based).
2. To introduce and extract texture-based feature i.e., GRanulometric Analysis (GRA).
3. To compute Similarity based statistical parameters.

4. Consideration of prominent static features e.g., string signature, byte-sequence, N-grams, OPCODE.
5. To introduce and apply SemiE based classifier.
6. To integrate feature and classifier to present proposed GRASE model which combines texture-based analysis with SemiE classifier for MD.

The above specified combination i.e., granulometric analysis with SemiE classifier, to the best of our knowledge, has not been assessed by the researchers.

This part provides the structured organization of the paper. Section II describes related work to the paper topic as well as GRA techniques used by other researchers in different domains. It also introduces the SemiE classifier. The section also explores varieties of techniques proposed by researchers which will help in formulating the problem statement. In Section III GRASE, the model of the proposed work, has been presented. Feature vector formation of granulometry feature and SemiE classifier mathematical model is described in this section. Section IV provides details about the experimental setup. Section V elaborates experimental results. The system performance of the proposed model is presented in Section VI. The overall conclusion is presented in Section VII.

II. Related Work

This section discusses various features and classification methods investigated by researchers in image-based techniques. It also describes the use of GRA for varieties of applications.

A. Image Based Method

Key benefits of representing malware executable as a 2D image are as follows. Firstly, once the similarity space has been formed, the data dimension does not affect processing. Secondly, it forms equally important clusters and finally, for the clear visualization one can display the similar clusters adjacent to each other [11] - [13].

Malware analysis using VAT with implementation of Self-Organizing Map (SOM) algorithm was proposed by Yoo [14]. S. Foresti [15] demonstrated usage of VAT to represent information like time (‘when’), IP address (‘where’), Data (‘what’) and estimated distances to other hosts. The first effort in the direction of visualization technique to visualize binary files for malware detection was done in 2008 [16]. Quist et al. used Ether Hypervisor framework to track and visually represent overall program flow by performing DA [17]. They named the DA framework VERA. Brute Force attack on Secure Shell (SSH) was identified using the VAT by Shiravi et al. [18] and N. Diakopoulos et al. [19]. They represented details of User IDs, Internet Protocol (IP) addresses and various anomalies with the help of different colours. Thus, large network packets were displayed using VAT with which security analysts were able to identify the minuscule details with the help of zoom option. Trinius et al. [20] proposed a novel concept of Malware Instruction SeT (MIST) for monitoring malware. They proposed the use of CW Sandbox for collecting information regarding performed actions and API calls. They used VAT to represent distance matrices of the features for the five malwares.

Further improvement in malware detection was observed when researchers started presenting binary executable sections as grayscale images. These images were used to present detailed structure of malware and even capable of showing small changes in the code. L. Nataraj et al. [21] put forth that the texture of a grey scale malware image can be used to identify similar patterns of the binary code.

Conti et al. [22] presented ‘Byte view’ visualization where each byte corresponds to a ‘single’ pixel of an image. The idea is feasible as image pixel and code byte value has the same range i.e., 00 to FF hexadecimal corresponding to different levels of gray scale. So, if the base malware code sequence is the same then it will produce similar images. They also introduced ‘Dot Plot’ visualization for comparing two images. This presentation helps to identify the presence of similar byte sequences. L. Nataraj et al. [21] visually observed that malware grey scale images were distinct for the different malware families and there was similarity in images for single malware families. Therefore, they extracted image texture-based features (GIST) and used KNN classifiers to classify different malware families. They achieved good average accuracy along with increased speed of malware detection. On a similar line Kancherla et al. [23] used byte plot (image of executable) and achieved 95% accuracy using SVM classifier. Vasanet. al. [24] classified malware images using Convolutional Neural Network (CNN) and Narayan et. al. [25] used Deep Neural Network (DNN).

To detect Trojan, Tian et al. [26] proposed Function Length Frequency (FLF) algorithm. Variable Length Instruction Sequences (VLIS) with ML for malware detection was proposed by Zolkipli [27]. Static Analyzer for Vicious Executable (SAVE) and Disassembled Code (MEDiC) were the two models suggested by Shankarapani et al. [28] for malware detection. The techniques were robust to code obfuscation; thus, results were promising.

Kong et al. [29] used L1 regularized technique to select the best feature from the set of features like PE header, disassembly code and n-gram. They evaluated system performance by using varieties of classifiers like KNN, SVM, Naïve Bayes (NB) and decision tree. They also figured out that PE header features are more prominent in MD.

Santos et al. [30] worked specifically on OPCODE. They tried to relate each OPCODE and calculated OPCODE sequence frequency. They used the same four classifiers used by Kong [29] to evaluate system performance.

Based on the work done by Trinius [20], shaid et al., [98] proposed DA based MDS by observing behavior of malware. They collected behavioral patterns for the operating system resources and API call sequences; and presented the same using color map. They found similarities between these color images using statistical methods. But collecting behavioral patterns is time consuming. K. han et al. [31] proposed hybrid MDS which extracts API calls and OPCODE sequences only. Image matrices were prepared from OPCODE sequences and further given to the classifier for the training purpose. Execution traces were extracted dynamically to avoid binary transformation strategies. However, the method was good for the small-scale MD.

For large-scale malware detection, researchers focused on similarity. According to [32]-[33] similarity should be calculated between all the pairs of points based on Euclidean distance. Maximum similarity corresponds to minimum distance. Normally, similarity patterns can be checked using 2D VAT like projection and semantic orientation [34]-[36]. Windows PE binary file was converted to grayscale image by Han et al. [20]. They calculated entropy of each and every row of an image using the Entropy Graph Generator (EGG). MD was performed based on similarity of the present file with the original binary file. Arefkhaniet et al. [37] proposed an image processing technique i.e., Local Sensitive Hashing for classifying similar malware images having high probability. Grey scale image was prepared by disassembling binary executable into OPCODE sequences [38]. They first reduced dimensionality using PCA and then used KNN to classify the malware images.

S. Rezaei [39] used similarity measurements by comparing OPCODE strings of malware files and tried to reduce detection time. Colored based VAT for analyzing malware attack chronology was used by Venkatraman [40], Zhang [41] and Wylie Shanks [42]. Successful system connection was demonstrated by them. J. Zhang et al. [9] extracted local texture features of grayscale image as well as OPCODE instructions of disassembly file, to train Random Forest

(RF) classification model. Shiqui et al. [43] extracted two features i.e., texture of malicious code and the frequency of instructions in the code. These features were used to train SoftMax classifiers and stacked auto encoders. Liu et al. [44] proposed enhancement of information density of malware images, where the '.text" section of malware code is visualized. To improve accuracy Wuechner et al. [45] proposed MDS based on data compression mining on the data flow graph.

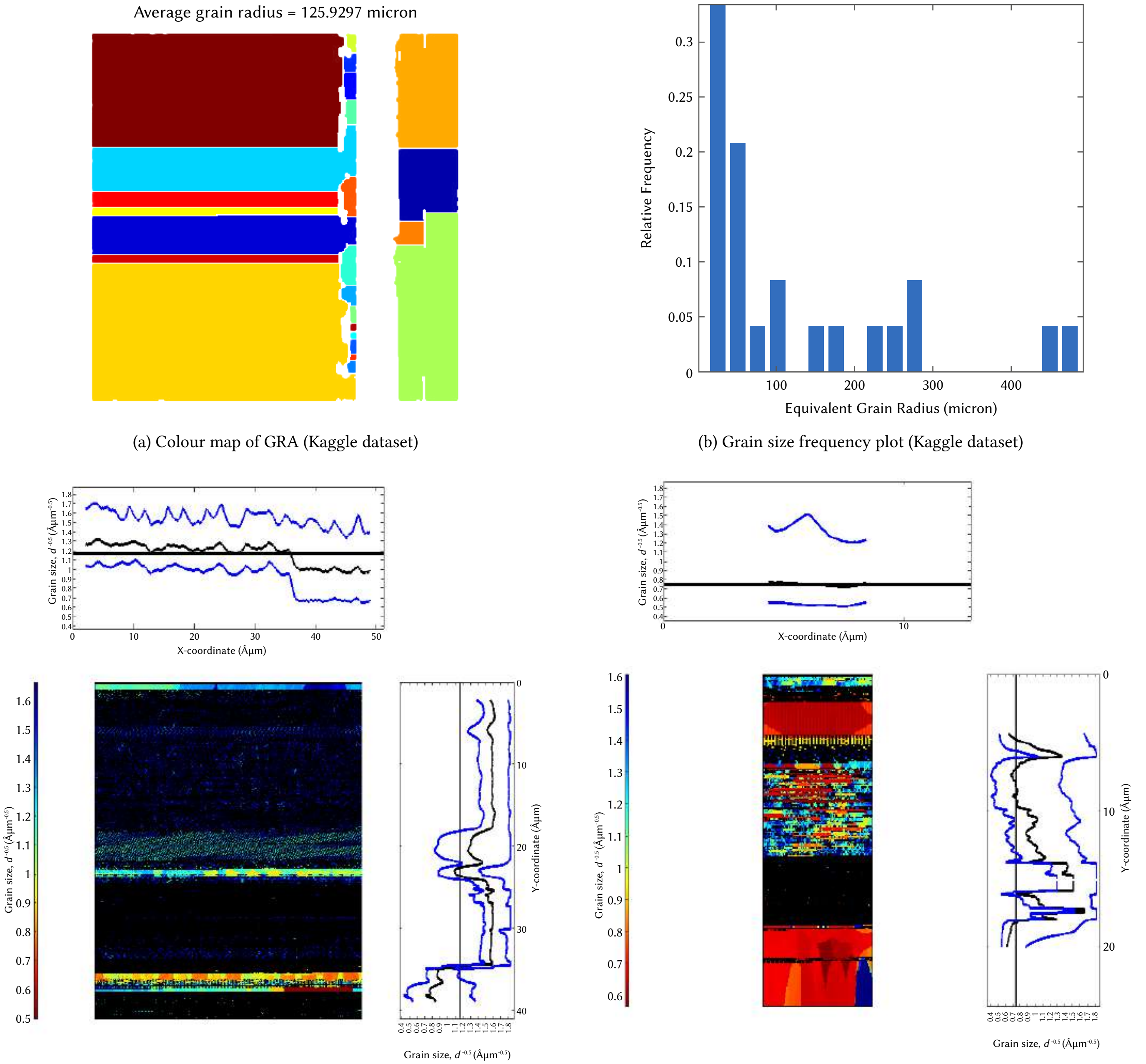
The overall extract from the above discussion is that the malware grey scale image analysis is not affected by code obfuscation, therefore opted by many researchers. In addition to that, texture is an important parameter for the grey scale image which can be used to find out similarity between malware images.

The next section provides related work done in Granulometry analysis and shows that the technique is versatile to image processing.

B. Granulometry Analysis

GRA concept was introduced in section I.B. This technique covers spatial as well as spectral characteristics of the malware grey scale image. GRA accuracy for classifying satellite images has been demonstrated by Kupidura et al. [5]. Basic GRA technique is based on morphological opening and closing. Extension of basic GRA is with a Multiple Structuring Element (MSE). Basic GRA and GRA with MSE have slightly different properties. So, it can give different results. GRA is also used to find the distribution of object sizes of an image [46]. We are proposing extraction of black patches from the grey scale image using successive closings by reconstruction.

GRA profiles (morphological profiles) can be used for image classification [47]. Thus, the technique is useful for malware family identification. Our main objective is to analyze the retrospective changes of the malware grey images for the family.



(a) Colour map of GRA (Kaggle dataset)

(b) Grain size frequency plot (Kaggle dataset)

(c) Malimg dataset - Adialer.C malware GRA [48]

(d) Malimg dataset - Dontovo. A malware GRA [48]

Fig. 1. GRA output for Kaggle Malware Family.

As GRA technique is relatively unknown, we are presenting two advantages of the system. The first advantage is multi-scality which is more suitable for malware detection. In this, we can obtain information about the texture grains of different sizes because of the possibility of successive application of increasing size of morphological opening and closing operations. Secondly, it is resistant to edge effects. In a typical texture analysis process, these edges have low texture, but still, it will provide high value of texture. This is due to the fact that those methods are based on spatial frequency analysis and imagery edges have high spatial frequency, resulting in high texture.

But as GRA is not based on this principle, it analyses value and number of the removed image elements, resulting in normal texture value of the edges in an image. Thus, the GRA seems to be more suitable for the MDS. Fig. 1 depicts granulometry output for one malware family from the Kaggle dataset and two malware families from the Malimg dataset. It shows a colour map for the different grain sizes. It also represents the relative frequency of different grain sizes present in an image. The grain size is measured using a point-sampled intercept length method. Related information is elaborated in section – III.A The next section introduces the main block of ML i.e., classifier.

C. Classifier

This section covers the classifier used by researchers for malware detection. Feature extractions and classification techniques are two basic blocks of MDS. API calls, system calls, n-gram and OPCODE are features that are still being used extensively in MDS. Work carried on by researchers implementing any approach is unique and makes use of different types of bytes and features related to hex. In addition to that we are also introducing a GRA feature. Thus, feature vectors extracted from the malware grey scale image have to be trained using the SemiE network. After training the SemiE model, system performance is evaluated by applying real time malware data. The next part elaborates upon the selection of SemiE.

The role of the classifier is crucial as it lays the groundwork for precision and accuracy. A research problem must cover the core subject matter and at the same time it must lead to hitherto undiscovered knowledge. This goal entails not only an extensive literature survey, but also mandates interpreting the surveyed information accurately to attain an appropriate research path. To make it more feasible to extract the requisite data, we have included graphical presentations (Refer Fig. 2).

The learning process is either Eager learning or Lazy learning (instance based). Key extract from the survey is both the techniques are at par. But the advancement in the Neural Network (NN) technique led to Deep Neural Network (DNN), which has opened up different segments altogether. But the major drawback of DNN is that the network is data hungry. This was the motivational factor where we thought of combining the advantages of lazy and eager and introduced Semi Eager learning. SemiE will reduce testing and training time, which is needed for malware detection methods, as it has to run in real time even as it detects malware in the fastest ways possible. This motivates us to choose the SemiE method with considerably low computational overhead and yet this technique has not yet been investigated for detection of malware. The next section elaborates the proposed model of GRASE.

III. Proposed Model: GRASE- GRanulometry Analysis With Semi Eager Classifier to Detect Malware

Fig. 3 shows the architecture of the GRASE model. The first step is to provide input to the model which is a malware file. There are three basic analysis techniques like SA, DA and hybrid, explored by researchers. While the SA method is the choice of many researchers, the hybrid technique is not so popular with them.

In step two, Malware Binary File (MBF) is read and features are extracted. The first feature set is SA based which has HEX dump-based features and disassembled file features. These features are common and must be used for malware prediction, therefore the same has been just specified and focus is given on the proposed technique.

The HEX dump-based features are n-gram, Meta-data (MD1), entropy [92] [93], Haralick and Local Binary Pattern (LBP) features. The disassembled file features are Meta-Data (MD2), symbol (SYM) [94], Register (REG) [97], Operation Code (OPCODE) [53], [95]-[96], DP and Section (SEC). Miscellaneous (MISC) feature should be done manually with the identification of keywords from the disassembled code. The Interactive Disassembler (IDA) tool can also be utilized for this purpose. Types of features that are extracted are number of imported DLLs, identifying strings viz. hkey_local_machine (it specifies access to specific paths of Windows registry), number of blocks in PE, etc. Hence, it is dependent on how experienced the MD software development engineer is.

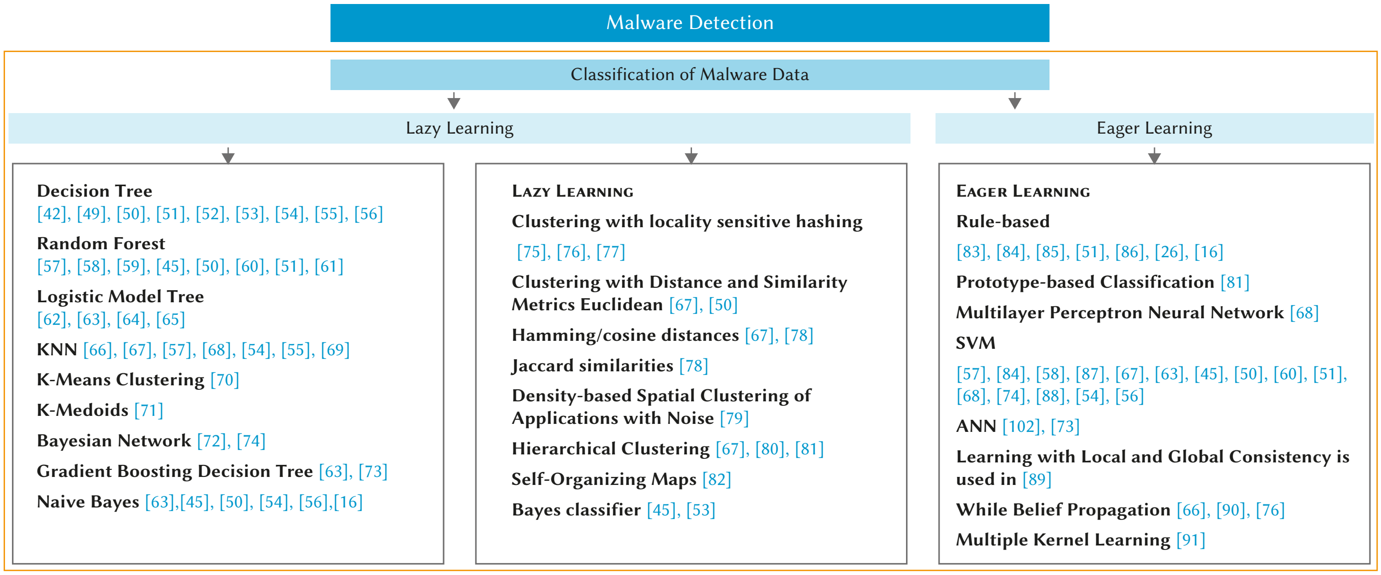

Malware Detection
Classification of Malware Data
Lazy Learning
Eager Learning
Decision Tree
[42], [49], [50], [51], [52], [53], [54], [55], [56]
Random Forest
[57], [58], [59], [45], [50], [60], [51], [61]
Logistic Model Tree
[62], [63], [64], [65]
KNN [66], [67], [57], [68], [54], [55], [69]
K-Means Clustering [70]
K-Medoids [71]
Bayesian Network [72], [74]
Gradient Boosting Decision Tree [63], [73]
Naive Bayes [63],[45], [50], [54], [56],[16]
Lazy Learning
Clustering with locality sensitive hashing
[75], [76], [77]
Clustering with Distance and Similarity Metrics Euclidean [67], [50]
Hamming/cosine distances [67], [78]
Jaccard similarities [78]
Density-based Spatial Clustering of Applications with Noise [79]
Hierarchical Clustering [67], [80], [81]
Self-Organizing Maps [82]
Bayes classifier [45], [53]
Eager Learning
Rule-based
[83], [84], [85], [51], [86], [26], [16]
Prototype-based Classification [81]
Multilayer Perceptron Neural Network [68]
SVM
[57], [84], [58], [87], [67], [63], [45], [50], [60], [51], [68], [74], [88], [54], [56]
ANN [102], [73]
Learning with Local and Global Consistency is used in [89]
While Belief Propagation [66], [90], [76]
Multiple Kernel Learning [91]

Fig. 2. Classifier based Literature Survey.

Very few malware programs make use of the packing technique and hence they do not use API calls. Instead, they contain a few OPCODEs. Generally, such programs use assembler related directives like Define Byte (db), Define Word (dw) and Define Double Word (dd). This feature plays a significant role in the classification of the varieties of malware families.

In step three, to extract GRanulometry features as well as Image Similarity based Statistical Parameter (ISSP), MBF is represented as a grayscale image. ISSP based features are Normalized Cross correlation (NCC), Average difference (AD), Maximum difference (MaxD), Singular Structural Similarity Index Module (SSIM), Laplacian Mean Square Error (LMSE), MSE and PSNR.

Sections III.A, III.B, III.C and III.D describe the proposed model shown in Fig. 3.

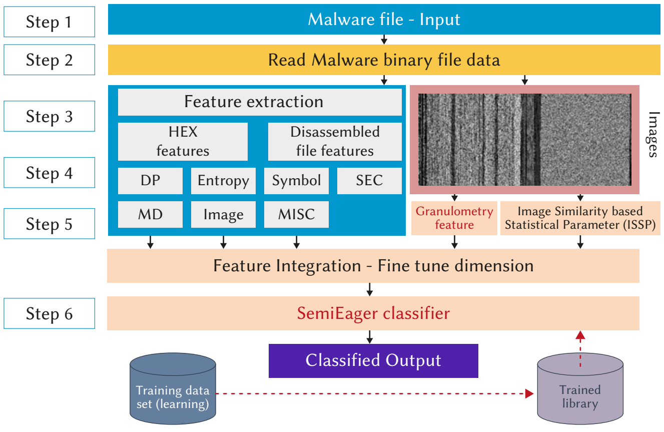


Fig. 3. Architecture Diagram: GRASE Model.

A. GRanulometry Analysis (GRA)

Step four is to generate granulometric profiles for each image pixel. GRA is based on the sequence of morphological opening and closing operations which are applied to gray scale image using set of known size and shape called the Structuring Element (SE). SE size is based on the pattern or a structure one would like to extract from the image. SE is normally a disk of size λ. In the process of closing by reconstruction, using SE will erase the dark spots of size less than λ during dilation process. Erased dark spots will not be recovered with multiple reconstructions, resulting in extraction of image structure having different sizes. Equation (1) represents granulometry density which describes the size of the image structures.

$$GRA_{\lambda} \;=\; \frac{|\emptyset_{\lambda}(I)-\emptyset_{\lambda-1}(I)|}{I}\;,\; \lambda \geq 1 \tag{1}$$

where, $\emptyset_{\lambda}\ (I)$ = *Closing by reconstruction*, λ − *Radius of disk* (*integer value*).

The operations are performed pixel wise. Granulometry profile may be written as

$$G(p) = \;[GRA_1(p), GRA_2(p), GRA_3(p) \ldots \ldots GRA_n(p)]$$

where, n = *Granulometrylevels. This parameter is configurable.*

GRA is used to measure the difference between two images by measuring the quantity of particles having different sizes by calculating Volume Weighted Average Grain Size (VWAGS). Refer Equation (2).

$$\text{VWAGS} = \;G_v \;=\; \frac{1}{V_I}\sum_{j=1}^{n} \quad V_j * G_j \tag{2}$$

Where, V_I *is total image size*, V_j *is the volume of grains corresponding to the grain size* G_j.

VWAGS will always be larger than the average grain size as per Eq. (1). It has been observed that VWAGS is able to capture the influence of grain size distribution [48].

Researches have used this technique to analyse satellite images. As in satellite images each pixel (granule) is important, on similar lines in malware image analysis each pixel carries important information of malware property. GRA can also be based on Multiple Structure Element (MSE). There are two main advantages of GRA. Firstly, it has a property of multi-scalability. Due to a greater number of morphological operations the information obtained will have texture grains of varieties of sizes.

Secondly, the analysis is resistant to edge effect. Edge effect is observed in fairly all texture analysis techniques where edges of the object normally have low texture, but it will get high value. This effect is observed as texture analysis methods are based on spatial frequency analysis and normally edges have a high spatial frequency, exhibiting high texture. GRA is not based on this fundamental as analysis is based on value and number of removed image elements and therefore edges are not displayed as areas of high texture. This property will help analyze malware images more accurately.

B. ISSP

Step five focuses on the similarity between two images. A comparison of malware images from the *'x'* family with themselves and the rest of the families is undertaken, and a similarity parameter matrix is computed based on this. There are two input images, namely, Reference image (R_I) and Input image (I_I). If R_I is from the *'x'* family then I_I represents the rest of the images from the *'x'* family and the images from the other families. During the entire process of computing the similarity parameters, R_I will remain constant. Since the number of images for every family is in the thousands their mean value will be computed [2].

The NCC method is utilized fortemplate matching. This is a procedure utilized for finding incidences of a pattern or object within an image. Eq. (3), is used to calculate NCC.

$$NCC(R_I, I_I) = C_{R_I I_I}\left(\widehat{R_I}, \widehat{I_I}\right) = \;\sum_{[m,n]\in R} \;\; \widehat{R_I}(m,n)\widehat{I_I}(m,n) \tag{3}$$

where, $\widehat{R_I} = \dfrac{R_I - \underline{R_I}}{\sqrt{\sum\left(R_I - \underline{R_I}\right)^2}}\;, \quad \widehat{I_I} = \dfrac{\left(I_I - \underline{I_I}\right)}{\sqrt{\sum\left(I_I - \underline{I_I}\right)^2}}$

AD provides the average [2] of change regarding the input image and the reference image. AD can be represented as follows:

$$AD(R_I, I_I) = \;\frac{1}{mn}\sum_{m=1}^{M} \; \sum_{n=1}^{N} \;[R_I(m,n) - I_I(m.n)] \tag{4}$$

MaxD provides the maximum of the error signal (i.e., the difference between the processed and reference image). *MaxD* is defined as follows:

$$MaxD(R_I, I_I) = \;\{[R_I(m,n) - I_I(m.n)\;]\} \tag{5}$$

SSIM is based on three factors [2], namely,luminance, contrast, and structure in order to be more in line with the workings of the human visual system. It is a perceptual metric that quantifies image quality degradation. This parameter is chosen as the malware developer alters the old code and comes up with the modified code. The modified code can be deemed to be the *'Noise'* element in an image. SSIM is defined as follows:

$$SSIM(R_I, I_I) = \;\left[l(R_I, I_I)^{\alpha}\;.\;c(R_I, I_I)^{\beta}\;.\,s(R_I, I_I)^{\gamma}\right] \tag{6}$$

where l = *luminance*, c = *contrast*, s = *structure*

The Laplacian error map [2] shows spatial error distribution across an image. The overall image quality is given by LMSE as follows:

$$LMSE(R_I, I_I) = \;\frac{\sum_{m=1}^{M} \;\; \sum_{n=1}^{N} \;\; [L(R_I(m,n)) - L(I_I(m.n)\;)]^2}{\sum_{m=1}^{M} \;\; \sum_{n=1}^{N} \;\; [L(R_I(m,n))]^2} \tag{7}$$

where $L((m, n))$ *is the Laplacian operator*

NAE measures [2] the numerical variance between the R_I and I_I. Additionally, the results that are closer to zero indicate that the image is highly similar to the original image and the results close to the value one mean that the quality of the image is very poor. NAE is calculated as follows:

$$NAE(R_I, I_I) = \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} |R_I(m,n) - I_I(m.n)|}{\sum_{m=1}^{M} \sum_{n=1}^{N} [R_I(m,n)]} \tag{8}$$

MSE and PSNR are used to compare the quality of image compression [2]. MSE represents the cumulative squared error between the R_I and I_I, whereas the PSNR represents a measure of the peak error. The lower the value of the MSE, the lower the error.

$$MSE(R_I, I_I) = \frac{1}{mn} \sum_{m=1}^{M} \sum_{n=1}^{N} [L(R_I(m,n)) - L(I_I(m.n))]^2 \tag{9}$$

$$PSNR = 10 * \frac{255^2}{MSE} \tag{10}$$

Once the ISSP has been computed, one more feature vector is generated which will be utilized to train the classifier.

To compute various statistical parameters, to begin with, all the malware families were segregated into different folders for the Kaggle dataset. There are 9 malware families in the Kaggle dataset so 9 folders were created. Grey images were produced after processing malware files. These images will be used to compute ISSP parameters. Malimg dataset is already organized and has grey scale images so it is ready to act as an input to the following algorithm. Table I presents the algorithm.

C. SemiE Classification Module

Finally step six corresponds to a SemiE classifier, whose mathematical model is explained here. In the training process, SemiE stores only the Centre Point for each class. SemiE training time complexity is $O(n)$, where n is the number of training instances. $O(k)$ is the prediction complexity of space and time, where k represents the number of categories.

1. SemiE algorithm

Let's say, $X =$ *Input space.*

It has *set of n dimension vector,* $X \subset R^n$.

$Y =$ *Output space.*

It has set of class labels $\{c_1, c_2, \dots \dots \dots, c_k\}$, *where* $c_i \in Z$

$P(X, Y) =$ *Joint probability distribution over X and Y*

Assume that, *Feature vector* $x \in X$ and *corresponding label* $y \in Y$.

$T\{(x_1, y_1), (x_2, y_2), \dots \dots \dots \dots, (x_N, y_N)\}$, is the training data set,

where x_i *and* y_i *are instances of X and Y, respectively*

$PS =$ *Partition of set* $T = \{(PS_1) \cup (PS_2) \cup \dots \dots \dots \dots \cup (PS_k)\}$,

Where $PS_j = \left\{(x_i, y_i)|_{y_i = c_j}\right\}$, *and* $\left(\frac{i}{j}\right) = 1,2, \dots \dots \dots, N$.

Learning algorithms learns these two probabilities, namely Conditional Probability Distribution and Prior Probability.

$$Prior\ Probability = P(Y = c_j), where\ j = 1,2, \dots., k$$

$$conditional\ probability\ distribution = P\left(X = x_i \ \middle|\ Y = c_j\right)$$

$$\therefore\ PosteriorProbability = P\left(Y = c_j \ \middle|\ X = x_i\right)$$

$$= \frac{P\left(X = x_i \ \middle|\ Y = c_j\right) * P(Y = c_j)}{\sum_{m=1}^{k} P\left(X = x_i \ \middle|\ Y = c_m\right) * P(Y = c_m)} \tag{11}$$

The algorithm, while predicting input x_i will provide the output class label based on maximum posterior probability.

TABLE I. Statistical Parameter Computation

```
Input   Folder structure is as follows.
        Main folder – contains sub folders equal to number of malware
        families ($i = 9$ for this case)
          - Sub-folders (9 malware families)
          - Each sub-folder has different number of images $j$
                        $R_I = Reference\ image$
                        $I_I = Input\ image$
        $i = Number\ of\ malware\ families\ in\ main\ directory\ (folder)$
        $j = Number\ of\ malware\ variants\ (images)\ of\ specific\ malware$
            $family\ in\ a\ subfolder$

        // Initialize empty array
        Parameter Array = {∅}
        for ($\beta = 0 ; \beta < i ; \beta++$)
        // Load reference image – first image of malware family
        $R_I = \beta[0]$
        for// select malware families one by one
            ($k = 0 ; k < i ; k++$)
        // Get number of images present of a specific malware family
          $j = size\ (k_{sub-folder})$
        for ($local_{cnt} = 0; local_{cnt} < j; local_{cnt}++$)
        // Load Input image from malware family
        $I_i = (k)[local_{cnt}]$
        // calculate SSIM
        $SSIM(R_i, I_i) = [l(R_i, I_i)^{\alpha} . c(R_i, I_i)^{\beta} . s(R_i, I_i)^{\gamma}]$
        where $l = luminance, c = contrast, s = structure$
        // Calculate MSE
        $MSE(R_I, I_I) = \frac{1}{mn} \sum_{m=1}^{M} \sum_{n=1}^{N} [L(R_I(m,n)) - L(I_I(m.n))]^2$
        // calculate PSNR
        $PSNR = 10 * \log_{10} \frac{255^2}{MSE}$
        // calculate Normalized Cross-Correlation (NK)
            $NCC(R_I, I_I) = C_{R_I I_I}(\widehat{R_I}, \widehat{I_I}) = \sum_{[m,n] \in R} \widehat{R_I}(m,n)\widehat{I_I}(m,n)$
        $\widehat{R_I} = \frac{R_I - \overline{R_I}}{\sqrt{\sum (R_I - \overline{R_I})^2}}, \widehat{I_I} = \frac{(I_I - \overline{I_I})}{\sqrt{\sum (I_I - \overline{I_I})^2}}$
        // calculate Normalized Absolute-Error (NAE)
        $NAE(R_I, I_I) = \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} |R_I(m,n) - I_I(m.n)|}{\sum_{m=1}^{M} \sum_{n=1}^{N} [R_I(m,n)]}$
        // calculate Maximum difference
        $MD(R_I, I_I) = \max\{[R_I(m,n) - I_I(m.n)]\}$
        // calculate Laplacian Mean Square Error (LMSE)
        $LMSE(R_I, I_I) = \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} [L(R_I(m,n)) - L(I_I(m.n))]^2}{\sum_{m=1}^{M} \sum_{n=1}^{N} [L(R_I(m,n))]^2}$
        where $L((m, n))$ is Laplacian operator
        // Store all the values in an array
        end
        // Take average of an array an obtain single value
        Parameter array(k)= [ mean (SSIM); mean (MSE); mean(PSNR);
        mean(NCC); mean(NAE);
        mean (MaxD); mean (LMSE)]
        end
        end
```

$$C = arg\ arg \frac{P\left(X = x_i \ \middle|\ Y = c_j\right) * P(Y = c_j)}{\sum_{m=1}^{k} P\left(X = x_i \ \middle|\ Y = c_m\right) * P(Y = c_m)}$$

As the denominator part is independent of c_j, it is constant, resulting in

$$C = arg\ argP\left(X = x_i \middle| Y = c_j\right) * P(Y = c_j) \tag{13}$$

2. Parameter Estimation

The Maximum Likelihood Estimation technique is used to estimate the parameters, therefore the indicator function and prior probability, both that may be computed as,

$$P\left(Y=c_j\right)=\frac{\sum_{i=1}^{N} \quad I(y_i=c_j)}{N} \tag{14}$$

$j = 1, 2, \dots \dots, k,\quad N = number\ of\ samples\ in\ trianing\ set$

$$I(u=v)=\begin{cases}1, & if\ u=v\\ 0, & otherwise\end{cases}$$

Posterior probability is calculated using Central Limit theorem. In this theorem whenever a number of samples N crosses a threshold limit T_H, then the input element x_i in the data set T generally follows a normal distribution having mean μ and variance σ^2.

Thus, if $|PS_j|$ is greater than T_H (let $T_H = 30$), then following the Central limit theorem x_i will have normal distribution with variance σ^2 and centred around μ_j.

∴ *conditional probability distribution* is given by,

$$P\left(X=x_i \;\middle|\; Y=c_j\right)=\frac{1}{\sqrt{2*\pi*\sigma^2}}e^{\left(-\frac{1}{2*\sigma^2}\left(x_i-\mu_j\right)^2\right)} \tag{15}$$

By substituting Eq. (14) to Eq. (12), we get

$$C = arg\ arg\frac{1}{\sqrt{2*\pi*\sigma^2}}e^{\left(-\frac{1}{2*\sigma^2}\left(x_i-\mu_j\right)^2\right)} \tag{16}$$

After discarding c_j independent term and constant values from the Eq. (15), we get,

$$C = arg\ arge^{\left(-\frac{1}{2*\sigma^2}\left(x_i-\mu_j\right)^2\right)*P(c_j)} \tag{17}$$

$$C = arg\ \mathrm{arg}\ \ ln\ e^{\left(-\frac{1}{2*\sigma^2}\left(x_i-\mu_j\right)^2\right)*P(c_j)} \tag{18}$$

$$C = arg\ arg\left(lnP\left(c_j\right)-\frac{1}{2*\sigma^2}\left(x_i-\mu_j\right)^2\right) \tag{19}$$

$$C = arg\ arg\left(\frac{1}{2*\sigma^2}\left(x_i-\mu_j\right)^2-ln\ lnP\left(c_j\right)\right) \tag{20}$$

$$C = arg\ arg\left(\left(x_i-\mu_j\right)^2-\ 2*\sigma^2*ln\ lnP\left(c_j\right)\right) \tag{21}$$

Condition – 1: If all the classes are having the same prior probabilities, then Eq. (21) can be written as

$$C = arg\ arg\left(\left(x_i-\mu_j\right)^2\right), if\ \ P(c_i)=P\left(c_j\right)(i\neq j) \tag{22}$$

where $\mu_j = j^{th}$ class centre

Classification of instance x_i, is based on the class centers of every class,

Let, $x_i = (x_i^1, x_i^2 \dots \dots \dots, x_i^n)$, then.

$$\mu_j=\frac{\sum_i I\left(y_i=c_j\right)x_i}{\sum_i I\left(y_i=c_j\right)}=\frac{\sum_i I\left(y_i=c_j\right)\left(x_i^1,x_i^2\dots\dots\dots,x_i^n\right)}{\sum_i I\left(y_i=c_j\right)} \tag{23}$$

Table II presents pseudo code of SemiE algorithm.

Significant properties of SemiE classifier are listed as follows. It performs incremental learning leading to training time complexity - $O(n)$ and prediction time complexity - $O(k)$. If the special case where the data point x_i is ready for classification but has equal distance from all the class labels is given, then the next level of decision making is done which is based on frequency. A regularization term will be assigned to the class label having the highest frequency.

TABLE II. SemiE Classifier Pseudo Code

```
Training Phase
Mean calculation, $\mu = \frac{\sum_{i=1}^{N} \quad x_i}{N}$                                  (24)
Variance calculation,
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu\ \ )^2}{N}$                                          (25)
$for\ j = 1\ to\ N$
    $for\ i = 1\ to\ k$
    $P\left(Y=c_j\right)=\frac{\sum_{i=1}^{N} I\left(y_i=c_j\right)}{N}$                     (26)
    $\mu_j \leftarrow \frac{\sum_i I\left(y_i=c_j\right)x_i}{\sum_i I\left(y_i=c_j\right)}=\frac{\sum_i I\left(y_i=c_j\right)\left(x_i^1,x_i^2\dots\dots\dots,x_i^n\right)}{\sum_i I\left(y_i=c_j\right)}$
$Regularization\ term\ r_j \leftarrow -\sigma^2*ln\ lnP\left(Y=c_j\right)$                     (27)
    $end$
$end$
Testing Phase
$Instance\ x_i\ is\ to\ be\ classified$
$C \leftarrow arg\ arg\left(\left(x_i-\mu_j\right)^2+r_j\right)$                               (28)
```

IV. Experimental Setup

The system performance is analyzed on a benchmark database from Kaggle[3] as well as the 'malimg' dataset[4]. Detailed information about both databases is given in Table III and Table IV. The major difference between these two datasets is that in 'malimg' set directly grey scale images were given, so, for the Kaggle dataset one additional coding function is required to convert malware files to grey scale image.

TABLE III. Kaggle and Malimg Datasets Basic Information

Header	Kaggle dataset	Malimg dataset
Download	Microsoft malware classification challenge from Kaggle	Vision research Lab
ID	Twenty-character hash value for unique identification of file	Thirty-two-character hash value for unique identification of file
Number of malware families / Size	9 / 0.5 Tera byte uncompressed	25 / 1.09 GB uncompressed, in image form
RAW data	HEX representation of the file's binary content	HEX representation of the file's binary content
Class	Integer representing malware family	Integer representing malware family
Metadata manifest	Log of various metadata information e.g. Function calls, Strings etc. extracted from the binary using IDA disassembler tool.	------------

V. Results

This section presents the results achieved throughout the process of malware analysis. Initially the Kaggle dataset malware files were converted to grey scale images, as shown in Fig. 4. It has been clearly noted that the image for each of the families is unique. 'malimg' dataset images are not shown as it is readily available from the website.

[3] https://www.kaggle.com/c/malware-classification/

[4] https://paperswithcode.com/dataset/malimg

TABLE IV. Kaggle and Malimg Datasets Description

Kaggle dataset			Malimg dataset					
Malware Family	**Malware category**	**Sample size**	**Malware Family**	**Malware category**	**Sample size**	**Malware Family**	**Malware category**	**Sample size**
Gatak	Backdoor	1013	Allaple.L	Worm	1591	Alueron.gen! J	Trojan	198
Obfuscator. ACY	obfuscated malware	1228	Allaple.A	Worm	2949	Malex.gen! J	Trojan	136
Kelihos_ver1	Backdoor	398	Yuner.A	Worm	800	Lolyda.AT	PWS	159
Tracur	TrojanDownloader	751	Lolyda.AA 1	PWS	213	Adialer.C	Dialer	125
Simda	Backdoor	42	Lolyda.AA 2	PWS	184	Wintrim.BX	Trojan Downloader	97
Vundo	Trojan	475	Lolyda.AA 3	PWS	123	Dialplatform.B	Dialer	177
Kelihos_ver3	Backdoor	2942	C2Lop.P	Trojan	146	Dontovo.A	Trojan Downloader	162
Lollipop	Adware	2478	C2Lop.gen! G	Trojan	200	Obfuscator.AD	Trojan Downloader	142
RAmnit	Worm	1541	Instantaccess	Dialer	431	Agent.FYI	Backdoor	116
			Swizzor.gen! I	Trojan Downloader	132	Autorun.K	Worm: AutoIT	106
			Swizzor.gen! E	Trojan Downloader	128	Rbot! gen	Backdoor	158
			VB.AT	Worm	408	Skintrim.N	Trojan	80
			Fakerean	Rogue	381			

TABLE V. Confusion Matrix

Malware	Malware Detection %								
	Ramnit	**Lollipop**	**Kelihos_ver3**	**Vundo**	**Simda**	**Tracur**	**Kelihos_ver1**	**Obfuscator. ACY**	**Gatak**
RAmnit	99.61	0.00	0.06	0.06	0.00	0.06	0.06	0.06	0.06
Lollipop	0.04	99.80	0.00	0.12	0.00	0.00	0.00	0.00	0.04
Kelihos_ver3	0.03	0.00	99.90	0.00	0.00	0.00	0.00	0.07	0.00
Vundo	0.00	0.21	0.21	99.37	0.21	0.00	0.00	0.00	0.00
Simda	0.00	0.00	0.00	0.00	95.24	2.38	2.38	0.00	0.00
Tracur	0.00	0.00	0.00	0.00	0.00	99.73	0.13	0.13	0.00
Kelihos_ver1	0.00	0.25	0.25	0.00	0.25	0.00	99.24	0.00	0.00
Obfuscator.ACY	0.08	0.08	0.08	0.08	0.00	0.00	0.08	99.43	0.16
Gatak	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.20	99.70

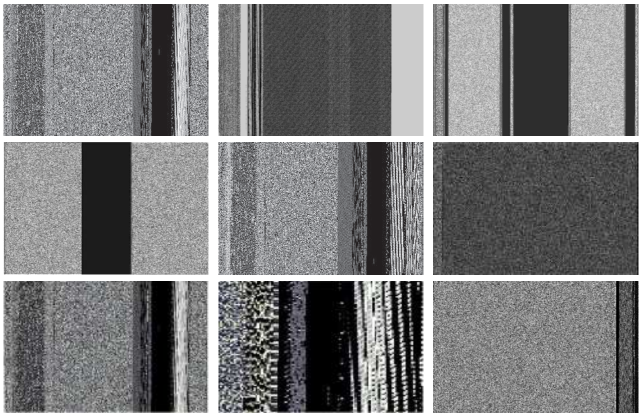


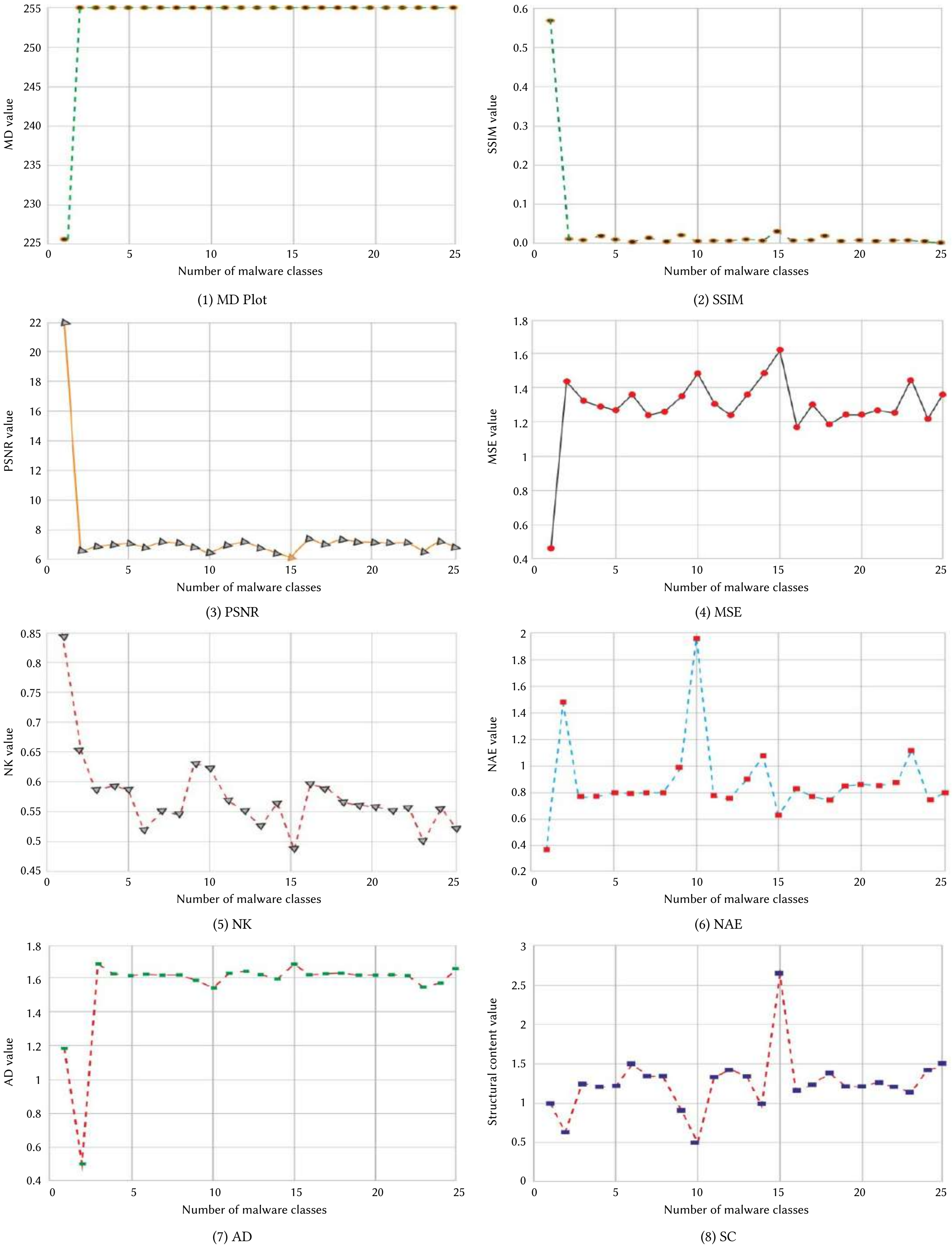
Fig. 4. Malware images of different malware families from the Kaggle dataset.

The next step was to extract the ISSP feature set from the grey scale images (refer section III.C). Suppose in the dataset, there are *MF number of malware families* For Kaggle dataset set *MF* is 9 and for Malimg dataset it is 25. There are *S number of samples* per malware family. For ISSP computation we require two images namely '*Reference Image* (R_I)' *and* '*Input image* (I_I)', R_I will remain constant through the iteration and I_I will change. Initially, R_I will be selected from the 1st malware family, and 1st image in the folder. Remaining all the images from 1st malware families and all the images of the remaining malware families will be now 'I_I'. R_I and I_I will be used to compute the ISSP parameters. Computed ISSP parameters per I_I will be stored in respective array say for example PSNR_array, NK_array and so on, with respect to malware family. After iterating through all the images of any one family, the mean value of an array will be computed. Hence, for every family there will be just one single mean value. The mean value matrix will be plotted. The same is illustrated in Fig. 5. Since the 'malimg' dataset is massive and has a large number of malware families we present the result of this dataset.

Fig. 5(1) shows the MD value. It is 225 for the first family of malware, but for the rest of the families the value is 255. Therefore, there is high structural similarity with self-family, but with other families a higher MD value reflects slower similarity. We can decide a threshold of 145 to 250 for differentiating between malware families. SSIM, PSNR, MSE, NK and NAE plots are on a similar line to MD i.e., clear bifurcation between self-family and other families, but the threshold values will be different. However, for the AD parameter for the self-family value is near to '0' and for other malware families it is either a high positive or high negative value (refer Fig. 5(7)), so AD demands for the hysteresis-based threshold. SC parameter value reflects overlap i.e., defining proper threshold is difficult, so we discarded this parameter for the training purpose (refer Fig. 5(8)). The remaining parameters can be used to train the SemiE classifier. The grain size selected for GRA analysis is 5,7,10 and 13 with MSE.

VI. Performance Evaluation

This section presents a confusion matrix and compares the results obtained from the proposed work with the state-of-the art methods.



(1) MD Plot

(2) SSIM

(3) PSNR

(4) MSE

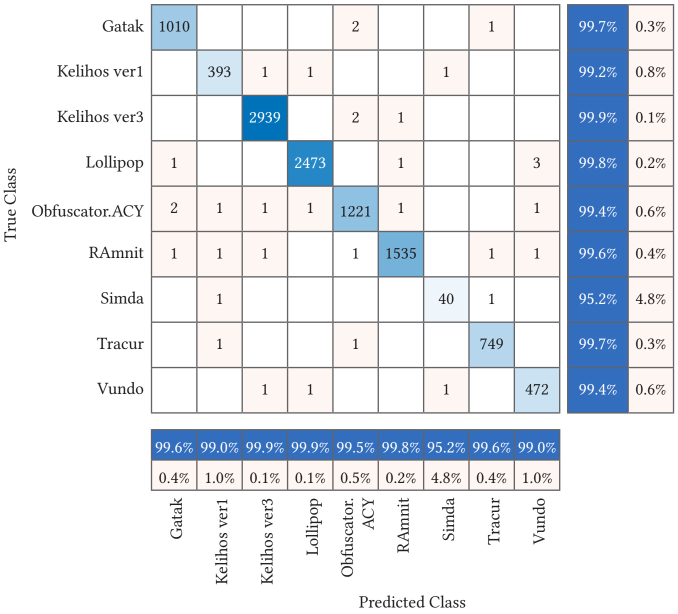
(5) NK

(6) NAE

(7) AD

(8) SC

Fig. 5. Parameter plots - (1) MD (2) SSIM (3) PSNR (4) MSE (5) NK (6) NAE (7) AD (8) SC.



	Gatak	Kelihos ver1	Kelihos ver3	Lollipop	Obfuscator.ACY	RAmnit	Simda	Tracur	Vundo		
Gatak	1010				2			1		99.7%	0.3%
Kelihos ver1		393	1	1			1			99.2%	0.8%
Kelihos ver3			2939		2	1				99.9%	0.1%
Lollipop	1			2473		1			3	99.8%	0.2%
Obfuscator.ACY	2	1	1	1	1221	1			1	99.4%	0.6%
RAmnit	1	1	1		1	1535		1	1	99.6%	0.4%
Simda		1					40	1		95.2%	4.8%
Tracur		1			1			749		99.7%	0.3%
Vundo			1	1			1		472	99.4%	0.6%
	99.6%	99.0%	99.9%	99.9%	99.5%	99.8%	95.2%	99.6%	99.0%		
	0.4%	1.0%	0.1%	0.1%	0.5%	0.2%	4.8%	0.4%	1.0%		

Fig. 6. Confusion Matrix Plot – Kaggle data set.

The proposed method was validated with the Big 2015 Kaggle and Malimg datasets and therefore results are compared with those research techniques that also have been validated with the same datasets.

Accuracy is a significant performance parameter for the MD system. It specifies accurate classification of malware. Accuracy is computed on the basis of the following equation:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} X\ 100$$

Table V shows the confusion matrix of the proposed MDS using Kaggle dataset.

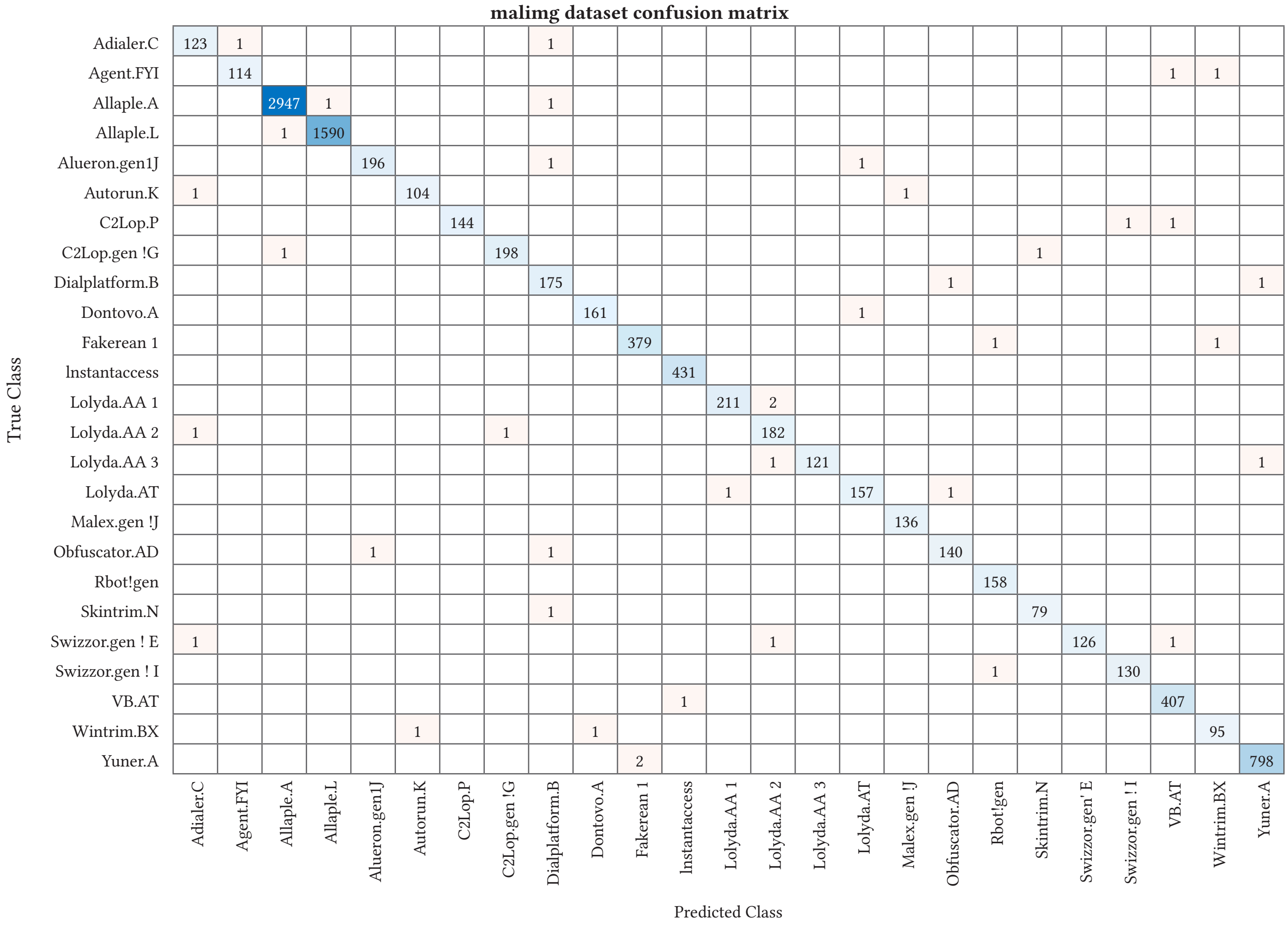
Graphical plots of confusion matrix are illustrated in Fig. 6 and Fig. 7.

A. State of the Art Comparison

As SemiE classifiers are not used by other researchers, we compared the results with learning algorithm presented by researchers. Table VI shows a comparative of the performance of the proposed and other methods.

Malimg and BIG 2015 datasets are the datasets commonly used by researchers. It has been observed that researchers work on either of the datasets, but this paper explores both the datasets. This makes the proposed work more robust.

CNN technique is also common in researchers. This technique is the basis for deep learning. Few researchers used variants of CNN i.e. DenseNet, ResNet-50 etc. Minimum accuracy of 98.23% to maximum 99.3% is achieved for malimg dataset. The proposed technique provides 99.03% accuracy. BIG2015 dataset is from Kaggle. Minimum 96.9% and maximum 99.73% accuracy achieved for the BIG2015 dataset. The proposed work provides 99.11% accuracy. Thus, the proposed provides minimum 99% accuracy for both the dataset.



	Adialer.C	Agent.FYI	Allaple.A	Allaple.L	Alueron.gen1J	Autorun.K	C2Lop.P	C2Lop.gen !G	Dialplatform.B	Dontovo.A	Fakerean 1	Instantaccess	Lolyda.AA 1	Lolyda.AA 2	Lolyda.AA 3	Lolyda.AT	Malex.gen !J	Obfuscator.AD	Rbot!gen	Skintrim.N	Swizzor.gen' E	Swizzor.gen ! I	VB.AT	Wintrim.BX	Yuner.A
Adialer.C	123	1							1																
Agent.FYI		114																					1	1	
Allaple.A			2947	1					1																
Allaple.L			1	1590																					
Alueron.gen1J					196				1							1									
Autorun.K	1					104											1								
C2Lop.P							144															1	1		
C2Lop.gen !G			1					198												1					
Dialplatform.B									175									1							1
Dontovo.A										161						1									
Fakerean 1											379								1					1	
Instantaccess												431													
Lolyda.AA 1													211	2											
Lolyda.AA 2	1							1						182											
Lolyda.AA 3														1	121										1
Lolyda.AT													1			157		1							
Malex.gen !J																	136								
Obfuscator.AD					1				1									140							
Rbot!gen																			158						
Skintrim.N									1											79					
Swizzor.gen ! E	1													1							126		1		
Swizzor.gen ! I																			1			130			
VB.AT												1											407		
Wintrim.BX						1				1														95	
Yuner.A											2														798

Fig. 7. Confusion Matrix Plot – Malimg data set.

TABLE VI. Proposed System Comparative Performance

Author	Dataset	Classifier	Accuracy (%)
J. Hemalatha et al., 2021 [99]	Malimg	Dense Net	98.23
V. Moussas et al., 2021 [100]	Malimg	Two level ANN	99.13
N. Marastoni et al., 2021 [101]	Malimg	CNN	98.5
M. Nisa et al., 2020 [102]	Malimg	CNN (AlexNet, Inception v3)	99.3
Ahmed Bensaoud et al, 2020 [103]	Malimg	Inception v3	99.24
Danish Vasan et al., 2020 [104]	Malimg	CNN	98.82
Hui Guo, et al., 2020 [105]	BIG 2015	ResNet50	99.73
Xianwei Gao et at., 2020 [106]	BIG2015	SSTL (Boost_ RNN)	96.9
S. A Roseline et al., 2020 [107]	BIG 2015	Deep random forest	97.2
Danish Vasan et al., 2020 [108]	ImageNet	CNN	98.82
Yuntao Zhao et al., 2020 [109]	BIG 2015	FRCNN with ImageNet	92.8
N. Bhodia et al., 2019[110]	malimg	DNN	94.8
Duc-Ly Vu et al.,2019 [111]	Author prepared dataset	Convolutional Transformation Network	99.14
Sung et al., 2021 [112]	BIG 2015	CNN	99.2
Proposed technique	**Kaggle**	**SemiE**	**99.11**
Proposed technique	**malimg**	**SemiE**	**99.03**

VII. Conclusion

Malware is a common attack on the internet. Developers of malware detection systems are continuously contending with cyber-attackers. For maintaining persistent pressure on cyber attackers, MDS developers should work out new strategies which can capture malware without any loss to the user and quarantine the same.

In this paper we have described a GRASE model which combines malware visualization, texture based GRanulometry (GRA) feature and Semi eager based classifier to classify malware images into different malware family classes.

System performance was evaluated using malimg and BIG-2015 Kaggle dataset. ‘malimg’ dataset is in the form of grayscale image, but BIG-2015 dataset required to be converted to gray scale images from the byte code of malware program. Each pixel in the gray scale malware image represents a code byte. Therefore, we applied a GRA technique where Granules’ or pixels are the important input to compute features.

With GRA additional features like n-gram, MD1, MD2, entropy, OPCODE, Register, symbols, data define, and Sections were used for generating feature vectors. This kind of learning approach is more suited to MDS because both, real time learning can be implemented with less time, and also testing or generating output in the form of malware detection is desideratum.

SemiE classifier with image-based visualization of feature vector resulted in an enhanced performance for classifying nine classes of malware and offered an overall accuracy of 99.11% with the Kaggle dataset and 99.03 % accuracy was achieved with the maling dataset (refer table VI).

Future scope for the proposed technique will focus on a diverse set of datasets to verify robustness of an algorithm. Presently a footprint of malware is available, but the objective of malware detection is customer security which should not be compromised. So, the task is to test a model if it can be used for predicting the new malware.

References

[1] G. Ekta, B. Divya, S. Sanjeev, “Malware Analysis and Classification: A Survey,” Journal of Information Security, vol. 5, no. 2, pp. 56-64, 2014.

[2] M. Deore, U. Kulkarni, “MDFRCNN: Malware Detection using Faster Region Proposals Convolution Neural Network,” International Journal of Interactive Multimedia and Artificial Intelligence, vol. 7, no. 4, pp. 146-162, 2022.

[3] E. Bou-Harb, M. Debbabi, and C. Assi, “Cyber Scanning: A Comprehensive Survey,” IEEE Communications Surveys & Tutorials, vol. 16, no. 3, Third Quarter, 2014.

[4] M. Dhalaria, E. Gandotra, “A Hybrid Approach for Android Malware Detection and Family Classification,” International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 6, pp. 174-188, 2021.

[5] P. Kupidura, “Wykorzystanie granulometrii obrazowej w klasyfikacji treści zdjęć satelitarnych,” Prace Naukowe Politechniki Warszawskiej. Geodezja, 2015.

[6] P. Kupidura, M. Skulimowska, “Morphological profile and granulometric maps in extraction of buildings in VHR satellite images,” Archives of Photogrammetry, Cartography and Remote Sensing, pp. 83–96, 2015.

[7] P. Kupidura, P. Koza, J. Marciniak, “Morfologia Matematyczna w Teledetekcji,” PWN: Warsaw, Poland, 2010.

[8] A. Haas, G. Matheron, J. Serra, “MorphologieMathématique et granulométriesen place. Ann. Des Mines,” vol. 12, pp. 768–782, 1967.

[9] E.R. Dougherty, J.B. Pelz, F. Sand, A, Lent, “Morphological Image Segmentation by Local Granulometric Size Distributions,” Journal of Electronic Imaging, vol. 1, pp. 46–60, 1992.

[10] A. Zakai and Y. Ritov, “Consistency and localizability,” *Journal of Machine Learning Research (JMLR)*, vol. 10, pp. 827–856, 2019.

[11] N. Cao and W. Cui, Introduction to Text Visualization, Atlantis Press, Paris, 2016.

[12] D. Keim, “Information visualization and visual data mining,” IEEE Transactions on Visualization and Computer Graphics, vol.8, no.1, pp.1–8, 2002.

[13] S. Few, “Information Dashboard Design - The Effective Visual Communication of Data, Sebastopol,” CA: O’Reilly, 2006.

[14] I. Yoo, “Visualizing windows executable viruses using self-organizing maps,” VizSEC/DMSEC ‘04: Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security, pp. 82-89. 10.1145/1029208.1029222, 2004.

[15] S. Foresti, J. Agutter, Y. Livnat, S. Moon, and R. Erbacher, “Visual correlation of network alerts,” IEEE Computer Graphics and Applications, vol. 26, no. 2, pp. 48–59,2006

[16] M. G. Schultz, E. Eskin, F. Zadok, S. J. Stolfo, “Data mining methods for detection of new malicious executables,” in: Proceedings 2001 IEEE Symposium on Security and Privacy. S&P 2001, Oakland, CA, USA, pp. 38-49, 2001.

[17] D. A. Quist and L. M. Liebrock, “Visualizing compiled executables for malware analysis,” 6th International Workshop on Visualization for Cyber Security, Atlantic City, NJ, pp. 27-32, 2009.

[18] H. Shiravi, A. Shiravi, and A. A. Ghorbani, “A survey of visualization systems for network security,” IEEE Transactions on Visualization and Computer Graphics, vol. 18, no. 8, pp. 1313–1329, 2012.

[19] N. Diakopoulos, D. Elgesem, A. Salway, A. Zhang, and K. Hofland, "Compare clouds: visualizing text corpora to compare media frames," in Proceedings of IUI Workshop on Visual Text Analytics, 2015.

[20] P. Trinius, T. Holz, J. Göbel and F. C. Freiling, "Visual analysis of malware behavior using tree maps and thread graphs," 6th International Workshop on Visualization for Cyber Security, Atlantic City, NJ, pp. 33-38, 2009.

[21] Nataraj, L., Karthikeyan, S., Jacob, G. and Manjunath, B, "Malware Images: Visualization and Automatic Classification," Proceedings of the 8th International Symposium on Visualization for Cyber Security, Article No. 4, 2011.

[22] Conti, G.; Bratus, S.; Shubina, A.; Lichtenberg, A.; Ragsdale, R.; Perez-Alemany, R.; angster, B.; Supan, M.A,"Visual Study of Binary Fragment Types," Black Hat: San Francisco, CA, USA, 2010.

[23] K. Kancherla and S. Mukkamala, "Image visualization-based malware detection," IEEE Symposium on Computational Intelligence in Cyber Security (CICS), Singapore, pp. 40-44, 2013.

[24] D. Vasan, M. Alazab, S. Wassan, H. Naeem, B. Safaei, Q. Zheng, "IMCFN: Image-based malware classification using fine-tuned convolutional neural network architecture,"Computer Networks, vol. 171, 107138, 2020.

[25] B.N. Narayanan, V.S.P Davuluru, "Ensemble Malware Classification System using Deep Neural Networks,"Electronics,vol. 9, no.5, 721, 2020.

[26] R. Tian, L. M. Batten, S. C. Versteeg, "Function length as a tool for malware classification," in: Malicious and Unwanted Software, MALWARE 2008. 3rd International Conference on, pp. 69-76,2008.

[27] Z. M. Fadli, A. Jantan, "An approach for malware behavior identification and classification," Computer Research and Development (ICCRD) 2011 3rd International Conference on, vol. 1, 2011.

[28] M. Shankarapani, S. Ramamoorthy, R Movva, S. Mukkamala, "Malware detection using assembly and api call sequences," Journal of Computing Virology, vol. 7, pp. 107-119, 2011.

[29] D. Kong, G. Yan, "Discriminant malware distance learning on structural information for automated malware classification," in: KDD '13: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, pp. 1357-1365, 2013.

[30] I. Santos, J. Devesa, F. Brezo, J. Nieves, P.G. Bringas, "OPEM: A Static-Dynamic Approach for Machine-Learning-Based Malware Detection," In: Herrero, Á., et al. International Joint Conference CISIS'12-ICEUTE´12-SOCO´12 Special Sessions. Advances in Intelligent Systems and Computing, vol 189. Springer, Berlin, Heidelberg, 2013.

[31] K. Han, J. H. Lim, and E. G. Im, "Malware analysis method using visualization of binary files," in Proceedings of the the2013 Research in Adaptive and Convergent Systems, pp. 317–321, Montreal, Quebec, Canada,2013.

[32] J. Jacobs and B. Rudis, "Data-driven security analysis, visualization, and dashboards," in Indianapolis, John Wiley & Sons, 2014.

[33] N. Cao, L. Lu, Y.-R. Lin, F. Wang, and Z. Wen, "Social Helix: visual analysis of sentiment divergence in social media," Journal of Visualization, vol.18, no.2, pp. 221–235, 2015.

[34] Dübel, Steve &Röhlig, Martin & Schumann, H. & Trapp, Matthias, "2D and 3D presentation of spatial data: A systematic review," 2014 IEEE VIS International Workshop on 3DVis (3DVis), Paris, France, 2014, pp. 11-18, doi: 10.1109/3DVis.2014.7160094.

[35] T. Songqing, "Imbalanced Malware Images Classification: a CNN based Approach," CoRR abs/1708.08042, 2017.

[36] W. B. Balakrishnan, "Security Data Visualization," SANS Institute Inc, 2014.

[37] M. Arefkhani and M. Soryani, "Malware clustering using image processing hashes," 9th Iranian Conference on Machine Vision and Image Processing (MVIP), Tehran, 2015, pp. 214-218, 2015.

[38] Q. Wu, Z. Qin, J. Zhang, H. Yin, G. Yang, K. Hu, "Android Malware Detection Using Local Binary Pattern and Principal Component Analysis," In: Zou B., Li M., Wang H., Song X., Xie W., Lu Z. (eds) Data Science. ICPCSEE 2017. Communications in Computer and Information Science, vol. 727, Springer, Singapore, 2017.

[39] S. Rezaei, A. Afraz, F. Rezaei, M. R. Shamani, "Malware detection using opcodes statistical features," in 2016 8th International Symposium on Telecommunications (IST), pp. 151–155, 2016.

[40] V. Sitalakshmi and M. Alazab, "Use of Data Visualization for Zero-Day Malware Detection,"Security and Communication Networks, vol. 2018, pp. 1728303:1-1728303:13, 2018.

[41] T. Y. Zhang, X. M. Wang Li, Z. Z. Li, F. Guo, Y. Ma, and W. Chen, "Survey of network anomaly visualization," Science China Information Sciences, vol. 60, no. 12, 121101, 2017.

[42] W. Shanks, "Enhancing Intrusion Analysis through Data Visualization," SANS Institute, Inc, 2015.

[43] L. Shiqi, T. Shengwei, S. Hua, and Y. Long, "Research on malicious code classification algorithm of stacked auto encoder," Application Research of Computers, vol. 35, no. 1, pp. 261–265, 2018.

[44] Y. Liu, Z. Wang, Y. Hou, and H. Yan, "Visualization and automatic classification of malicious code with enhanced information density,'' *J. Tsinghua Univ. (Natural Sci. Ed.)*, vol. 59, no. 1, pp. 914, 2019.

[45] T. Wuchner, M. Ochoa, A. Pretschner, "Robust and effective malware detection through quantitative data flow graph metrics," in: Detection of Intrusions and Malware, and Vulnerability Assessment, Springer, pp. 98-118, 2015.

[46] R.W. Conners, C.A. Harlow, "A Theoretical Comparison of Texture Algorithms," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-2, no. 3, pp. 204-222, 1980.

[47] S.G. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, no. 7, pp. 674-693, 1989.

[48] Lehto P, Romanoff J, Remes H, Sarikka T. "Characterisation of local grain size variation of welded structural steel," Weld World, vol. 60, pp. 673 678, 2016, http://dx.doi.org/10.1007/s40194-016-0318-8

[49] S. Srakaew, W. Piyanuntcharatsr, S. Adulkasem, "On the comparison of malware detection methods using data mining with two feature sets," Journal of Security and Its Applications, vol. 9, no. 3, pp. 293-318, 2015.

[50] D. Uppal, R. Sinha, V. Mehra, V. Jain, "Malware detection and classification based on extraction of api sequences," in: ICACCI, IEEE, pp. 2337-2342, 2014.

[51] R. Islam, R. Tian, L. M. Batten, S. Versteeg, "Classification of malware based on integrated static and dynamic features," Journal of Network and Computer Applications, vol. 36, no.2, pp. 646-656, 2013.

[52] S. Nari, A. A. Ghorbani, "Automated malware classification based on network behavior," in: Computing, Networking and Communications (ICNC), 2013 International Conference on, IEEE, pp. 642-647, 2013.

[53] I. Santos, F. Brezo, X. Ugarte-Pedrero, P. G. Bringas, "Opcode sequences as representation of executables for data-mining-based unknown malware detection," Information Sciences, vol. 231, pp. 64-82, 2013.

[54] I. Firdausi, C. Lim, A. Erwin, A. S. Nugroho, "Analysis of machine learning techniques used in behavior-based malware detection," in: ACT '10, IEEE, pp. 201-203, 2010.

[55] F. Ahmed, H. Hameed, M. Z. Shafiq, M. Farooq, "Using spatio-temporal information in api calls with machine learning algorithms for malware detection," in: Proceedings of the 2nd ACM workshop on Security and artificial intelligence, ACM, pp. 55-62, 2009.

[56] J. Z. Kolter, M. A. Maloof, "Learning to detect and classify malicious executables in the wild," Journal of Machine Learning Research, pp. 2721-2744, no. 7, pp. 2721-2744, 2006.

[57] M. Ahmadi, G. Giacinto, D. Ulyanov, S. Semenov, M. Tromov, "Novel feature extraction, selection and fusion for effective malware family classification," CoRR abs/1511.04317, 2016

[58] B. J. Kwon, J. Mondal, J. Jang, L. Bilge, T. Dumitras, "The dropper effect: Insights into malware distribution with downloader graph analytics," in: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, ACM, pp. 1118-1129, 2015.

[59] W. Mao, Z. Cai, D. Towsley, X. Guan, "Probabilistic inference on integrity for access behavior-based malware detection," in: International Workshop on Recent Advances in Intrusion Detection, Springer, pp. 155-176, 2015.

[60] P. M. Comar, L. Liu, S. Saha, P. N. Tan, A. Nucci, "Combining supervised and unsupervised learning for zero-day malware detection," in: INFOCOM, 2013 Proceedings IEEE, pp. 2022-2030, 2013.

[61] M. Siddiqui, M. C. Wang, J. Lee, "Detecting internet worms using data mining techniques," Journal of Systemic, Cybernetics and Informatics, vol. 6, no. 6, pp. 48-53, 2009.

[62] M. Graziano, D. Canali, L. Bilge, A. Lanzi, D. Balzarotti, "Needles in a haystack: Mining information from public dynamic analysis sandboxes for malware intelligence," in: USENIX Security '15, pp. 1057-1072, 2015.

[63] J. Sexton, C. Storlie, B. Anderson, "Subroutine based detection of APT

malware," Journal of Computer Virology and Hacking Techniques, vol. 12, pp. 225-233, 2016.

[64] G. E. Dahl, J. W. Stokes, L. Deng, D. Yu, "Large-scale malware classification using random projections and neural networks," in: Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 3422-3426, 2013.

[65] S. Palahan, D. Babi_c, S. Chaudhuri, D. Kifer, "Extraction of statistically significant malware behaviors," in: Computer Security Applications Conference, ACM, pp. 69-78, 2013.

[66] E. Raff, C. Nicholas, "An alternative to ncd for large sequences," lempel-zivjaccard distance, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 1007-1015, 2017.

[67] A. Mohaisen, O. Alrawi, M. Mohaisen, "Amal: High-fidelity, "behavior based automated malware analysis and classification," Computers & Security, vol. 52, pp. 251-266, 2015.

[68] D. Kong, G. Yan, "Discriminant malware distance learning on structural information for automated malware classification," in: ACM SIGKDD '13, nKDD '13, ACM, New York, NY, USA, pp. 1357-1365, 2013.

[69] T. Lee, J. J. Mody, "Behavioral classification," in: EICAR Conference, pp. 1-17, 2006.

[70] K. Huang, Y. Ye, Q. Jiang, "Ismcs: an intelligent instruction sequence-based malware categorization system," in: Anti-counterfeiting, Security, and Identification in Communication, 2009, IEEE, pp. 509-512, 2009.

[71] Y. Ye, T. Li, Y. Chen, Q. Jiang, "Automatic malware categorization using cluster ensemble," in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 95- 104, 2010.

[72] M. Eskandari, Z. Khorshidpour, S. Hashemi, "Hdm-analyser: a hybrid analysis approach based on data mining techniques for malware detection," Journal of Computer Virology and Hacking Techniques, vol. 9, pp. 77-93, 2013.

[73] Z. Chen, M. Roussopoulos, Z. Liang, Y. Zhang, Z. Chen, A. Delis, "Malware characteristics and threats on the internet ecosystem," Journal of Systems and Software, vol. 85, pp. 1650-1672, 2012.

[74] J. Yonts, "Attributes of malicious files," Tech. rep., The SANS Institute, 2012.

[75] J. Upchurch, X. Zhou, "Variant: a malware similarity testing framework," in: 2015 10th International Conference on Malicious and Unwanted Software (MALWARE), IEEE, pp. 31-39,2015.

[76] A. Tamersoy, K. Roundy, D. H. Chau, "Guilt by association: large scale malware detection by mining file-relation graphs," in: Proceedings of the 20th ACM SIGKDD, ACM, pp. 1524-1533, 2014.

[77] U. Bayer, P. M. Comparetti, C. Hlauschek, C. Kruegel, E. Kirda, Scalable, "behavior-based malware clustering," in: NDSS, Vol. 9, pp. 8-11, 2009.

[78] M. Polino, A. Scorti, F. Maggi, S. Zanero, "Jackdaw: Towards Automatic Reverse Engineering of Large Datasets of Binaries," in: Detection of Intrusions and Malware, and Vulnerability Assessment, Lecture Notes in Computer Science, Springer International Publishing, pp. 121-143, 2015.

[79] P. Vadrevu, B. Rahbarinia, R. Perdisci, K. Li, M. Antonakakis, "Measuring and detecting malware downloads in live network traffic," in: Computer Security ESORICS 2013: 18th European Symposium on Research in Computer Security, Egham, UK, September 9-13, 2013. Proceedings, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 556-573, 2013.

[80] J. Jang, D. Brumley, S. Venkataraman, "Bitshred: feature hashing malware for scalable triage and semantic analysis," in: Computer and communications security, ACM, pp. 309-320, 2011.

[81] K. Rieck, P. Trinius, C. Willems, T. Holz, "Automatic analysis of malware behavior using machine learning," Journal of Computer Security, vol. 19, no. 4, pp. 639-668, 2011.

[82] S. Attaluri, S. McGhee, M. Stamp, "Profile hidden Markova models and metamorphic virus detection," Journal in Computer Virology, vol. 5, pp. 151-169, 2009.

[83] G. Liang, J. Pang, C. Dai, "A behavior-based malware variant classification technique," International Journal of Information and Education Technology, vol. 6, no. 4., pp. 291-295, 2016.

[84] Z. Feng, S. Xiong, D. Cao, X. Deng, X. Wang, Y. Yang, X. Zhou, Y. Huang, G. Wu, "Hrs.: A hybrid framework for malware detection," in: Proceedings of the 2015 ACM International Workshop on Security and Privacy Analytics, ACM, pp. 19-26, 2015.

[85] M. Ghiasi, A. Sami, Z. Salehi, "Dynamic VSA: a framework for malware detection based on register contents," Engineering Applications of Artificial Intelligence, vol. 44, pp. 111- 122, 2015.

[86] M. Lindorfer, C. Kolbitsch, P. M. Comparetti, "Detecting environment sensitive malware," in: Recent Advances in Intrusion Detection, Springer, pp. 338-357, 2011.

[87] C.T. Lin, N.J. Wang, H. Xiao, C. Eckert, "Feature selection and extraction for malware classification," Journal of Information Science and Engineering, vol. 31, no. 3, pp. 965-992, 2015.

[88] B. Anderson, D. Quist, J. Neil, C. Storlie, T. Lane, "Graph-based malware detection using dynamic analysis," Journal in Computer Virology, vol. 7, no. 4, pp. 247-258, 2011.

[89] I. Santos, J. Nieves, P. G. Bringas, "Ch. Semi-supervised Learning for Unknown Malware Detection," International Symposium on Distributed Computing and Artificial Intelligence, Springer Berlin Heidelberg Berlin, Heidelberg, pp. 415-422, 2011.

[90] D. H. Chau, C. Nachenberg, J. Wilhelm, A. Wright, C. Faloutsos, "Polonium: Tera-scale graph mining for malware detection," in: ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 131-142, 2010.

[91] B. Anderson, C. Storlie, T. Lane, "Improving malware classification: bridging the static/dynamic gap," in: Proceedings of the 5th ACM workshop on Security and artificial intelligence, ACM, pp. 3-14, 2012.

[92] D. Baysa, R. Low, and M. "Stamp. Structural entropy and metamorphic malware," Journal of Computer Virology and Hacking Techniques, vol. 9, no. 4, pp. 179–192, 2013.

[93] R. Lyda and J. Hamrock, "Using entropy analysis to find encrypted and packed malware," IEEE Security and Privacy, vol. 5, no. 2, pp. 40–45, 2007.

[94] A. Moser, C. Kruegel, and E. Kirda "Limits of static analysis for malware detection," In Computer Security Applications Conference, 2007. ACSAC 2007, Twenty-Third Annual, pp. 421–430, 2007.

[95] D. Bilar. "Statistical structures: Finger printing malware for classification and analysis," InBlackhat, 2006.

[96] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. "Evasion attacks against machine learning at test time," in H. Blockeel, K. Kersting, S. Nijssen, and F. Železný (editors), Machine Learning and Knowledge Discovery in Databases, vol. 8190 of Lecture Notes in Computer Science, pp. 387–402, Springer Berlin Heidelberg, 2013.

[97] M. Christodorescu, S. Jha, S. Seshia, D. Song, and R. Bryant. "Semantics-aware malware detection," In Security and Privacy, 2005 IEEE Symposium on, pp. 32–46, 2005.

[98] S.Z.M. Shaid, M.A. Maarof, "Malware behaviour visualization," Jurnal Teknologi, vol. 70, no. 5, 2014.

[99] J. Hemalatha, S.A. Roseline, S. Geetha, S. Kadry, R. Damaševičius, "An Efficient DenseNet-Based Deep Learning Model for Malware Detection," Entropy, vol. 23, no. 3, 344, 2021.

[100] V. Moussas, A. Andreatos, "Malware Detection Based on Code Visualization and Two-Level Classification," Information, vol. 12, no. 3, 118, 2021, https://doi.org/10.3390/info12030118

[101] N. Marastoni, R. Giacobazzi, M. Dalla Reda, "Data augmentation and transfer learning to classify malware images in a deep learning context," Journal of Computer Virology Hacking Techniques, vol. 17, no. 279-297, 2021.

[102] M. Nisa, J.H. Shah, S. Kanwal, M. Raza, M.A. Khan, R. Damaševičius, T. Blažauskas, "Hybrid Malware Classification Method Using Segmentation-Based Fractal Texture Analysis and Deep Convolution Neural Network Features," Applied Sciences, vol. 10, no. 14, 4966, 2020.

[103] A. Bensaoud, N. Abudawaood, J. Kalita, "Classifying Malware Images with Convolutional Neural Network Models," ArXiv, abs/2010.16108, 2020.

[104] D. Vasan, M. Alazab, S. Assan, H. Naeem, B. Safaei, Q. Zheng, "IMCFN: Image-based malware classification using fine-tuned convolutional neural network architecture," Computer Networks, Vol. 171, 107138, 2020.

[105] H. Guo, S. Huang, C. Huang, F. Shi, M. Zhang, Z. Pan, "Binary File's Visualization and Entropy Features Analysis Combined with Multiple Deep Learning Networks for Malware Classification," Security and Communication Networks, vol. 20, 8881760, 2020.

[106] X. Gao, C. Hu, C. Shan, B. Liu, Z. Niu, H. Xie, "Malware classification for the cloud via semi-supervised transfer learning," Journal of Information

Security and Applications, vol. 55, 102661, 2020.

[107] S. A. Roseline, S. Geetha, S. Kadry and Y. Nam, "Intelligent Vision-Based Malware Detection and Classification Using Deep Random Forest Paradigm," in IEEE Access, vol. 8, pp. 206303-206324, 2020.

[108] D. Vasan, M. Alazab, S. Wassan, B. Safaei, Q. Zheng, "Image-Based malware classification using ensemble of CNN architectures (IMCEC)," Computers & Security, Vol. 92, 101748, 2020.

[109] Y. Zhao, W. Cui, S. Geng, B. Bo, Y. Feng and W. Zhang, "A Malware Detection Method of Code Texture Visualization Based on an Improved Faster RCNN Combining Transfer Learning," IEEE Access, vol. 8, pp. 166630-166641, 2020.

[110] N. Bhodia, P. Prajapati, F. Di Troia, M. Stamp, "Transfer Learning for Image-Based Malware Classification," doi: 10.5220/0007701407190726, 2019.

[111] D. -L. Vu, T. -K. Nguyen, T. V. Nguyen, T. N. Nguyen, F. Massacci and P. H. Phung, "A Convolutional Transformation Network for Malware Classification," 2019 6th NAFOSTED Conference on Information and Computer Science (NICS), Hanoi, Vietnam, 2019, pp. 234-239.

[112] K.-S. Sung, W. Na, "A study on the implementation of a system providing reliable malware information service," International Journal of Electrical Engineering & Education, vol. 58, no. 2, pp. 517-530, 2021.

Mahendra Deore

M. Deoreis working as an Asst. Professor in Computer Engineering Department atMKSSS's Cummins College of Engineering for Women, Pune 411051, India. He was awarded his Master of Technology Degree from Bharati Vidyapeeth Deemed University College of Engineering, Dhankawadi, Pune. He received doctoral degree from Swami Ramanand Teertha Marathwada University, Nanded, Indiain 20022. His areas of interest are big data, Security, Computer Networks and Machine learning. He has Fourteen years' experience in teaching.

Manoj Tarambale

He received graduate degree (B.E.) in Electrical Engineering from University of Pune (SPPU), India in 1992, post graduate degree (M.E.) in Control System from Shivaji University, Kolhapur, India in 2002 and completed research work (PhD) in Electrical – Biomedical Image Processing from PACIFIC University, Udaipur, India in 2018. He has one-year industrial experience and thirty years teaching experience. At present, he is Associate Professor of electrical engineering department and Principal of PVG's College of Engineering and Technology & G K Pate Institute of Management, Pune-09, India. His main research interests are control system engineering, electrical vehicle technology, robotics & automation, bio-medical image processing, electronic instrumentation, and medical diagnosis (AI, ML & DS based).

Jambi Ratna Raja Kumar

Prof (Dr) Ratna Raja Kumar Jambi Completed his PhD (CSE) degree from Maharishi University of Information Technology, Lucknow, Uttar Pradesh in 2019, Master of Technology (CSE) from Pondicherry Central University in 2007. He has 17 years of Teaching and Research Work. He is having Patents at National, international level and has published Papers in artificial Intelligence and Machine Learning. He has received the award as "Innovative Leader" form World Education Summit & Awards in 2019 at New Delhi.

Sachin Sakhare

Dr. Sachin R. Sakhare is working as a Professor and Head of the Computer Engineering Department at Vishwakarma Institute of Information Technology, Pune, India. He has 27 Years of experience in engineering education. He is recognized as PhD guide by Savitribai Phule Pune University and currently guiding 8 PhD scholars. He is a life member of CSI, ISTE and IAEngg. He has Published 51 research communications in national, international journals and conferences, with around 393 citations and H-index 7. He has authored 6 books which is published by Springer Nature, CRC Press and IGI Global. He worked as a reviewer of journals published by Elsevier, Wiley, Hindawi, Springer, Inder science, and IETE. He worked as a reviewer for various conferences organized by IEEE, Springer, and ACM. He worked as a member of the Technical and Advisory Committees for various international conferences. Dr. Sachin has Delivered invited talks at various Conferences, FDP's and STTP's as well as to PG and PhD students. He has guided 26 PG students. He has filed and published 07 patents out of which 01 Indian, 03 Australian and 02 south African patents are granted.

Chatbot-Based Learning Platform for SQL Training

Antonio Balderas*, Rubén Baena-Pérez, Tatiana Person, José Miguel Mota, Iván Ruiz-Rube

Departamento de Ingeniería Informática, Universidad de Cádiz, Escuela Superior de Ingeniería, Puerto Real (Spain)

* Corresponding author. antonio.balderas@uca.es

Received 19 July 2021 | Accepted 1 April 2022 | Early Access 3 May 2022

Abstract

Learning the SQL language for working with relational databases is a fundamental subject for future computer engineers. However, in distance learning contexts or unexpected situations like the COVID-19 pandemic, where students had to follow lectures remotely, they may find it hard to learn. Chatbots are software applications that aim to have conversations with people to help them solve problems or provide support in a specific domain. This paper proposes a chatbot-based learning platform to assist students in learning SQL. A case study has been conducted to evaluate the proposal, with undergraduate computer engineering students using the learning platform to perform SQL queries while being assisted by the chatbot. The results show evidence that students who used the chatbot performed better on the final SQL exam than those who did not. In addition, the research shows positive evidence of the benefits of using such learning platforms to support SQL teaching and learning for both students and lecturers: students use a platform that helps them self-regulate their learning process, while lecturers get interesting metrics on student performance.

Keywords

Chatbot, E-Learning, Interactive Learning Environments, SQL, Technology-Enhanced Learning.

DOI: 10.9781/ijimai.2022.05.003

I. Introduction

The application of computer technologies supports lectures to design more engaging and effective learning experiences for their students to help them achieve learning goals [1]. Furthermore, in contexts where lecturers are not always available, as may be the case in distance education or under conditions of confinement as experienced in the recent pandemic caused by COVID-19, education becomes more dependent on technology, and its availability [2], [3].

After being successfully used in other fields, a technological artefact that is beginning to be incorporated into e-learning platforms is the conversational agent, also known as chatbot [4],[5]. Chatbots interact with users using natural language, answer users' questions, provide them with personalised answers [6], and even manage emotions in the dialogue [7]. There are two kinds of chatbots: task-oriented chatbots designed to support a specific task or context and non-task-oriented chatbots designed to emulate casual conversations with the human user [8]. Leading companies have developed several chatbots for industry and research: Apple Siri, Microsoft Cortana, Facebook M, Google Assistant and IBM Watson Assistant.

This research study has its origin in the learning of the Structured Query Language (SQL). SQL is a specific domain language designed to manage and retrieve information from Relational Database Management Systems (RDBMS) [9]. Within the SQL, there is a set of instructions for managing tables and views known as Data Manipulation Language (DML). DML is based on relational algebra and relational tuple calculation. Consequently, not having a consolidated knowledge of these principles can severely disadvantage students when learning SQL in-depth [10].

This study proposes the use of a task-oriented chatbot to support students in learning SQL. The chatbot guides students in solving various SQL queries through natural language interac tions, motivated by their questions and attempts. Besides, the chatbot provides different levels of difficulty, adapting to each learner's specific needs and pace and being available at any moment. Finally, we made it available to students enrolled in a university course on Databases to evaluate the chatbot.

The rest of the paper is organised as follows. The second section reviews various software systems supporting the learning of SQL. In the third section, the developed chatbot is described. The fourth section presents the experience carried out with students. The fifth section presents and discusses the results. Finally, conclusions and future work are shown.

II. State of the Art

Databases are one of the fundamental subjects for all computer engineering students. In a 2018 report on the 54 Database courses in Computer Engineering degrees at Spanish universities [11], it is shown that conceptual design is taught in 86% of them, the relational model in 81%, relational algebra and calculus in 55%and finally, logical design in 74%. In addition, students learn and practice SQL through a RDBMS such as Oracle, MySQL and SQLServer.

Please cite this article as:
A. Balderas, R. Baena-Pérez, T. Person, J. M. Mota, I. Ruiz-Rube, "Chatbot-Based Learning Platform for SQL Training", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 6, pp. 135-145, 2024, http://dx.doi.org/10.9781/ijimai.2022.05.003

Although the most straightforward SQL queries are usually easy to learn for students, learning SQL in depth is often not an easy task for them if they do not have a good foundation in algebra and relational computing. The academy has tried to improve the situation by using interactive tools to help students learn SQL [12]. To this end, in the following subsection, several tools that have been recommended by the teaching staff and used by students are described. In the second subsection, different uses of chatbots in learning contexts are presented.

A. Tools for Learning SQL

The first learning tool analysed is SQL Course[1]. SQL Course is an interactive online tool that provides a theoretical and practical course in SQL, including the basic theory of the different clauses, and an online interpreter to write solutions to the proposed exercises. Although it uses a friendly approach, students' experience with this website is highly improvable in two respects. Firstly, each query produces a reload on the page when doing the exercises, creating unnecessary waiting times. Secondly, there is no interaction with the tool when solving the exercises: if a query is correct, the results are displayed, and if it is wrong, a generic error message is thrown.

SQLzoo is another interactive web tool for learning SQL [13]. This tool is similar to SQL Course, although it partially improves the feedback the student receives. Once the student proposes a solution to an exercise, the application returns the result of the execution of the query and indicates whether or not the result is as expected.

In line with the tools mentioned above, SQL easy[2] is another web tool that provides feedback on the accuracy of the queries provided to the proposed exercises.However, unlike the previous ones, it offers a more friendly design, although it does not guide and support students in the development of the queries beyond informing them if the query result is correct or not.

Khanacademy[3] is a more up-to-date and comprehensive platform than those previously presented. It provides similar functionality and a follow-up through interactive videos that guide students in trying out the different queries. Although the video feature is interesting, it uses a generic approach that is not adapted to the needs of each student. On the other hand, it does not provide more help than checking whether the proposed query's result is expected when solving exercises.

The customisation of learning tools to enable self-regulation of student learning is of interest to lecturers [14]. In this vein, SQL-Knot was integrated with WebEx system and SQL-Lab to build a user model for each student, which was used to adapt some of its components based on individual progress [15]. SQL-KnoT is a tool that generates problems that students must solve through SQL queries. SQL-Lab is a tool that supports the formulation of SQL queries and the testing of their results.

B. Chatbots in Education

Chatbots, also known as conversational agents, are computer programs designed to converse with human persons, employing Natural Language Processing (NLP) and sentiment analysis techniques [16].

The evolution of chatbots can be differentiated into three phases [17]. The first phase took place during the 1970s and 1980s when the study of natural language interfaces commenced [18]. The first chatbot ELIZA is created in this phase based on pattern matching and a template-based response selection scheme [19]. However, it is limited by the inability to save a record of the conversation [20].

The second phase happened when the internet became increasingly popular, allowing users to have online conversations with many people. One of the most significant advances was the creation of ALICE in 1995, developed with a new Artificial Intelligence Mark-up Language (AIML), which allows a stimulus-response template to be defined [21].

The third and current phase is when significant advances in NLP and Automatic Speech Recognition (ASR) technologies have been achieved. As a result, a large number of new alternatives are created, such as Apple Siri [22], IBM Watson Assistant [23] or Google Assistant [24], among many others. They stand out for providing a degree of naturalness and speed in their responses that sometimes lead the user to doubt if the interlocutor is a human person [25].

Due to the level of sophistication reached, chatbots are increasingly adopted in different fields such as commerce [26], entertainment [27], health [28] or the public sector [29]. In the educational field, chatbots have an essential potential due to two factors; their capacity to communicate through natural language and the possibility of offering support to students at any moment. These two factors allow the human teacher to delegate mechanical or repetitive tasks to the chatbot and assume other tasks with a higher cognitive level [17]. The use of chatbots by students helps to improve the learning process and teaching due to the information they gather during their use [30]. Although chatbots can be confused with Intelligent Tutoring Systems (ITS), these systems are different. ITS are domain-specific, and their interaction is strongly conditioned by following a series of steps, while chatbots base their interaction on natural language conversation.

Education-focused chatbots can be classified into two types [17]: tutors who support the learning process and exercise and practice chatbots for skill acquisition. Tutors are teaching agents who work as learning partners, providing dialogue, collaboration and reflection. In contrast, exercise and practice chatbots are based on the presence of stimulus-cognition-response and reinforcement.

Within the first of these types of chatbots, there is Genie [31]. Designed by Deakin University using IBM Watson, this chatbot is part of a strategic plan for the digitalisation of the university, and it aims to resolve any doubts students may have regarding the campus. Another chatbot is Jill Watson [32]. Also developed with IBM Watson, this chatbot answers students' questions and can raise doubts during the conversation. A study of 300 students did not detect that it was a chatbot, and it was even nominated as the best teacher of the year. Furthermore, Prof. Watson is a conversational agent for primary school students learning to program [33]. The students who used it stated that thanks to this tool, they learned in a dynamic and fun way while consolidating aspects of programming.

Concerning the second type of these education-oriented chatbots, those oriented towards language learning stand out. Duolingo [34] is a popular tool for language learning, supporting 34 different languages to choose from. This tool presents a learning mechanism where a series of questions are formulated, and users must answer them in writing or using their voice.

III. Chatbot for Learning SQL

This section describes a chatbot for learning SQL and is divided into five subsections. The first subsection presents the platform architecture. The second one details its development, the third one describes the chatbot implementation, the fourth its functionality, and the fifth one explains how the interaction between the student and the chatbot is performed.

[1] https://www.sqlcourse.com/

[2] https://www.sql-easy.com/

[3] https://es.khanacademy.org/computing/computer-programming/sql

A. Architecture

Fig. 1 shows the platform's architecture as a container diagram as proposed in the C4 model [35].

The SQL learning platform is presented as a web application to which both students (1) and lecturers (2) connect using a username and a password. The web application is based on the Laravel framework. It was chosen due to its extensive documentation, clarity and performance in database queries thanks to Eloquent Object-Relational Mapping (ORM) [36].

The web application is connected to two databases. Firstly, the web application is connected (3) with a database that stores the platform's configuration: users, SQL statements for the exercises and levels of difficulty.

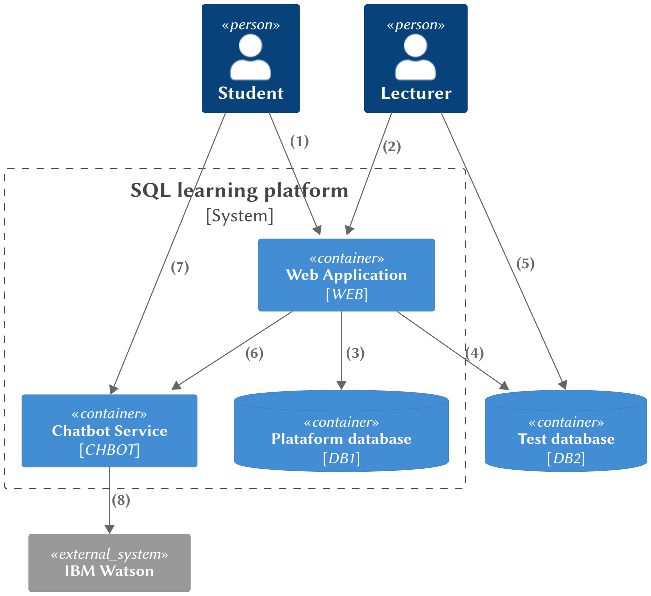


Fig. 1. SQL learning platform architecture.

Secondly, the web application is connected with a database (4) that contains the tables on which the queries issued by the students to solve the proposed exercises will be run. Lecturers can connect to this database (5) to insert or delete data records or incorporate new data structures.

These databases are currently hosted by Amazon Web Services (AWS). The choice of AWS was based on the reliability provided by this platform, its ease of administering the services contracted, and the non-functional aspects it offers, such as security [37].

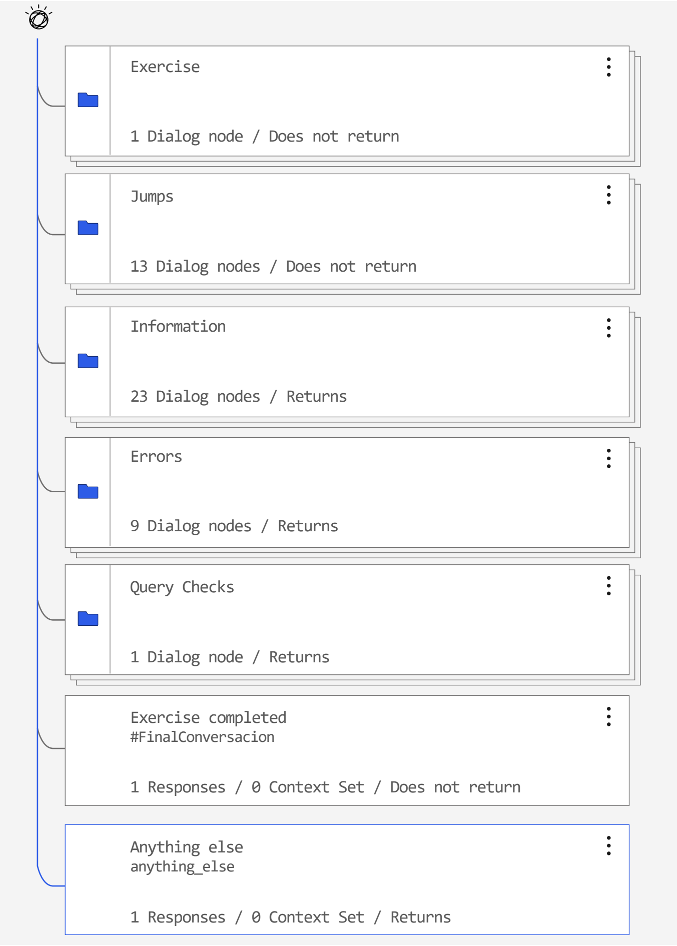
Finally, to implement the chatbot functionality, the web application connects to IBM Watson (8) via a service API (6). It was decided to use a Node.js server, which allows concurrent queries, to connect users with IBM Watson artificial intelligence [38]. Students will interact with the chatbot for support in resolving queries (7).

B. Chatbot Development

An NLP platform is required to design the artificial intelligence responsible for providing the chatbot with the ability to interpret natural language. IBM Watson was chosen because of its monitoring functions and integration with Big Data tools and its widespread use in education research [33]. The conversation flow between the user and the chatbot is defined using a series of nodes interconnected. These nodes indicate to the chatbot how it should operate when a particular event occurs during the conversation. These events, known as intents, are defined by a series of phrases instructing the chatbot about what text input is expected from the user. Within these nodes, intents are sometimes accompanied by entities used to identify specific content in the users' phrases.

Fig. 2 shows how the nodes have been grouped according to their function in the conversation. The group of nodes called ”Exercise” determines the steps of the exercise and the corresponding help for each step. These nodes are nested among themselves to follow a consecutive order in the execution of the exercise.

The group of nodes called ”Jumps” leads the flow of the conversation according to the achievement of the exercise by the user. This achievement level ranges from a wrong exercise to a partially or correct one. In order to resolve any doubts about the SQL syntax, the ”Information” node group has been created. Thanks to this group, the user obtains an extensive description of the concepts presented in the exercises and examples of their use.



Exercise
1 Dialog node / Does not return

Jumps
13 Dialog nodes / Does not return

Information
23 Dialog nodes / Returns

Errors
9 Dialog nodes / Returns

Query Checks
1 Dialog node / Returns

Exercise completed
#FinalConversacion
1 Responses / 0 Context Set / Does not return

Anything else
anything_else
1 Responses / 0 Context Set / Returns

Fig. 2. IBM Watson Dialog Nodes.

The last two of these groups, ”Errors” and ”Query Checks”, express the interactions with the server. ”Query Checks” conveys to the server the SQL queries issued by the user, while ”Errors” optionally sends to the user explanations of the results obtained during the exercises. Finally, the second to last node informs users that the current exercise has ended, while the last one indicates that the conversational agent has not understood their question.

C. Chatbot Dialogue Implementation

This subsection describes how the conversation is implemented in the IBM Watson Assistant.

To identify which of the different dialogue nodes a conversation should be directed to, the assistant will try to correspond the message received with one of the nodes through the conditions. Fig. 3 shows how the dialogue nodes work. For defining the con-ditions, the intentions and context variables are used. The list of intents defined in the implementation of this assistant is as follows:

- #Acknowledgements
- #Help
- #Goodbye
- #Start
- #End_Conversation
- #Get_Information
- #Go_Next
- #What_you_can_do
- #Regular
- #Hello
- #Syntax_error
- #Start_exercise
- #Finished

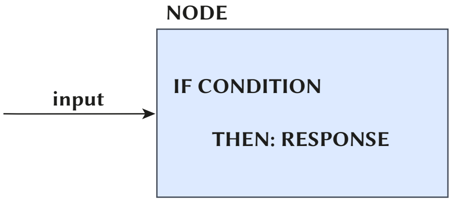
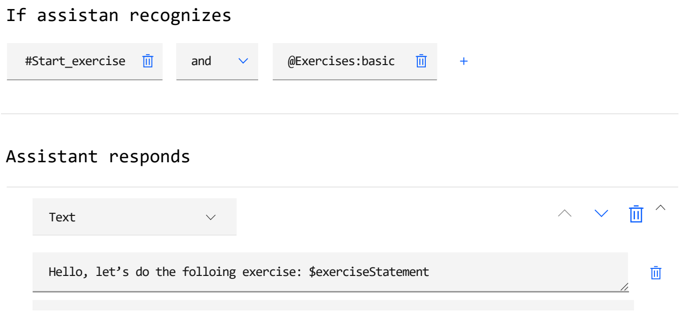


Fig. 3. Operation of IBM Watson nodes.

Below, we describe the intention #Start_exercise. In order for the assistant to identify that the student wants to start the exercise, training sentences must be given in advance. In this case, the following sentences were used for this intent:

- I want to start the exercise
- I want to start
- Let's start the exercise

Thus, if the assistant identifies a sentence similar to one of these, it will perform the actions specified in this node. You can see this example in Fig. 4. The assistant will reply with the specified message if it recognises the intent to start the exercise (#Start_exercise) and receives an exercise entity (@Exercise) of the basic type, provided through a context variable.



If assistan recognizes

#Start_exercise and @Exercises:basic

Assistant responds

Text

Hello, let's do the folloing exercise: $exerciseStatement

Fig. 4. Response exercise configuration.

A context variable is a variable that is defined at a given moment in the conversation, which is always available to the assistant and whose value can be changed throughout the conversation. In this way the conversation can be personalised. In addition, the assistant can collect information that can then be referred to or reused later in the conversation.For example, in the SQL chatbot, the context variables used are two:

- Context variable to store the exercise statement. This is defined at the start of the exercise and is not modified, being available in case the student asks the assistant for it at a given moment in the conversation.
- Context variable to know in which step the student is. As the student progresses through a query, this context variable is updated. For example, if the student asks for help right at the start, the chatbot tells him to start by identifying which table he needs (with the ”show tables” query). On the other hand, if the student has almost completed the query but is failing to implement ”HAVING”, he/she will receive help with the HAVING clause.

In the example shown in Fig. 4, the statement is in the context variable $exerciseStatement, which is used to indicate the statement to the student.

Fig. 5 shows the whole process of integration and communication of the learning platform and the chatbot with the IBM Watson Assistant. When a student proceeds to initiate an exercise on the platform, the exercise data is requested from the view through its identifier (1). This request is processed by the controller, which consults the exercise data from the exercise database (2) and returns the exercise information (3) so that the view can show it to the student on the screen (4). The controller also sends a conversation start message to the chatbot (5), which establishes communication and initiates a session with the IBM Watson Assistant (6), defining the context variable and returning a message (7) to display a welcome message and a message with the statement on the chat dialog box.

Now the student will be able to interact with the platform by typing a SQL query in the console or asking the chatbot for help through the chatbot dialogue box.

- If the student types a SQL query in the console (8), the controller submits the query (9) to the DB server and collects the results (10). The controller returns these results to the view (11) to be displayed to the student and sends them to the chatbot (12). The chatbot sends a message to IBM Watson with context variable and the result of the query (13), and IBM Watson returns an appropriate response to the result received by the query (14).
- If the student asks for help directly to the chatbot through the dialogue box (15). The chatbot sends the message and the context variable to the assistant (16). The IBM Watson assistant returns a response appropriate to the help requested and the point of the exercise the student is going through (17).

D. Functionality Description

The SQL learning platform that integrates the chatbot offers two different user profiles: lecturer and student. As a student, the platform allows performing SQL exercises, classified into three levels of difficulty, which will be enabled as they complete them. There is positive evidence of the benefits of offering flexibility to students in managing their learning [39]. Thanks to the levels of difficulty, each learner can learn at his or her own pace.

For each exercise, the SQL learning platform displays the problem statement (see Fig. 6, top centre box). On a text entry box just below the statement box, the student can write and issue the queries and observe the results. On the right side of the platform, a panel is displayed from which the user will interact with the chatbot. This chatbot will provide clues on how to solve the exercises and clear and concise instructions on the syntax and semantics of the SQL language.

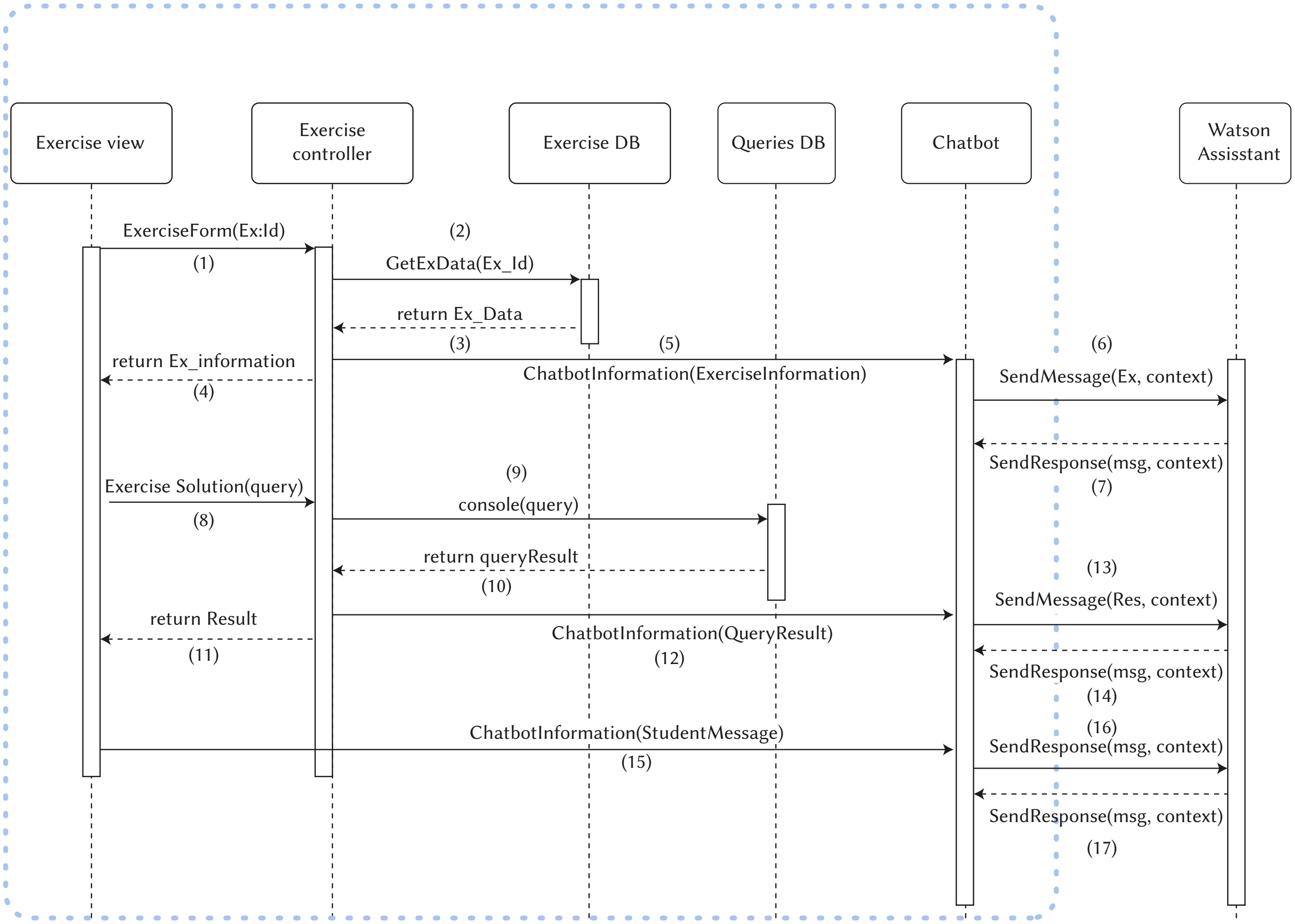
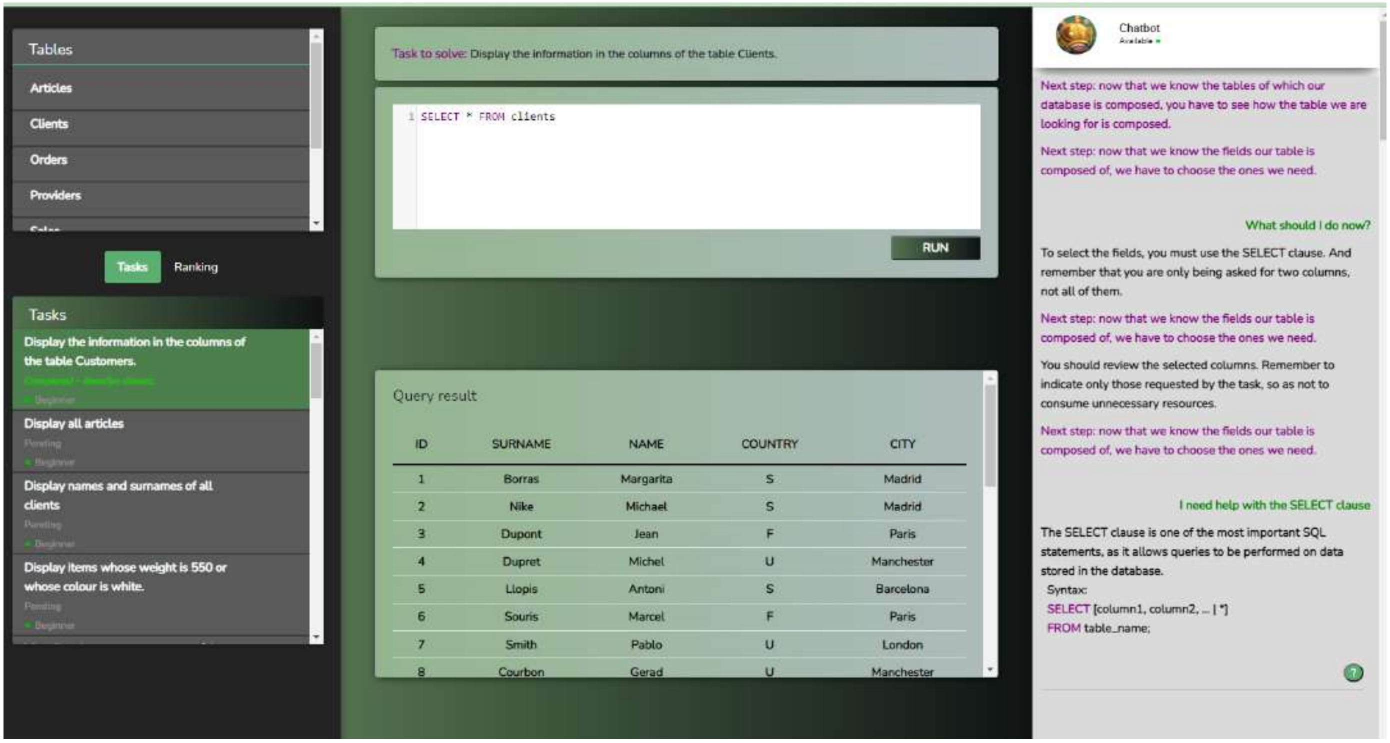


Fig. 5. Design and integration diagram of components that are part of the learning platform.



Display the information in the columns of the table Customers.

Display all articles
Pending
Beginner

Display names and surnames of all clients
Pending
Beginner

Display items whose weight is 550 or whose colour is white.
Pending
Beginner

Task to solve: Display the information in the columns of the table Clients.

```
SELECT * FROM clients
```

Query result

ID	SURNAME	NAME	COUNTRY	CITY
1	Borras	Margarita	S	Madrid
2	Nike	Michael	S	Madrid
3	Dupont	Jean	F	Paris
4	Dupret	Michel	U	Manchester
5	Llopis	Antoni	S	Barcelona
6	Souris	Marcel	F	Paris
7	Smith	Pablo	U	London
8	Courbon	Gerad	U	Manchester

Next step: now that we know the tables of which our database is composed, you have to see how the table we are looking for is composed.

Next step: now that we know the fields our table is composed of, we have to choose the ones we need.

What should I do now?

To select the fields, you must use the SELECT clause. And remember that you are only being asked for two columns, not all of them.

Next step: now that we know the fields our table is composed of, we have to choose the ones we need.

You should review the selected columns. Remember to indicate only those requested by the task, so as not to consume unnecessary resources.

Next step: now that we know the fields our table is composed of, we have to choose the ones we need.

I need help with the SELECT clause

The SELECT clause is one of the most important SQL statements, as it allows queries to be performed on data stored in the database.

```
 Syntax:
 SELECT [column1, column2, ... | *]
 FROM table_name;
```

Fig. 6. The main screen of the SQL learning platform where the student performs the exercises while interacting with the chatbot.

The learning platform also includes a ranking page showing the users who have completed the exercises, providing a gamified experience. Gamification of learning activities increases higher motivation, thus, positively affecting student academic performance [40].

The SQL learning platform enables lecturers to include new exercises by adding their problem statements, an SQL query to generate the correct solution, and additional clues to help them. In addition, the platform provides a screen from which to access the set of exercises and conversations held between students and the chatbot, which can be used to conduct further learning analytics.

E. Student-Chatbot Interaction

Student interaction with the chatbot involves a series of steps that are described below:

1. Exercise: the SQL learning platform provides the student with the statement of the exercise to be solved.
2. Hint: the chatbot provides a hint for the student to take the next step in the exercise.
3. Propose a solution or ask for help: the student can introduce a query to solve the exercise or ask for help from the chatbot (in case of asking for help, the student would go back to step 2).
4. Feedback: after entering the SQL query, it is evaluated by the RDBMS, and the result returned is displayed on the screen.
   - If the query result is correct, the exercise is considered solved, and the student proceeds to the next exercise.
   - Otherwise, the chatbot informs the student that the query defined is incorrect, provides a new hint and goes back to step 3.

IV. Evaluation

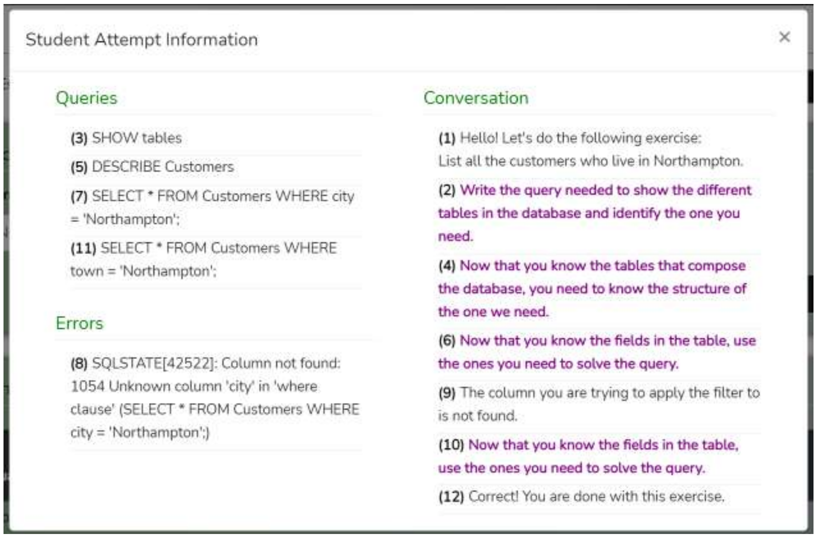
The SQL learning platform has been evaluated in a case study conducted with students of the subject Databases, compulsory in the second year of the Degree in Computer Engineering at the University of Cadiz (Spain) during part of the 2019/2020 academic year. The objective of this preliminary evaluation, in which 59 students participated, was to inspect the chatbot's behaviour with real students to assess the quality and effectiveness of their responses. In addition, the SQL learning platform served as reinforcement material for preparing the SQL final examination of the course, especially suitable for this course due to the suspension of face-to-face classes because of the COVID-19 pandemic.

The Database course consists of 3 components: theory, practice and assignments. The assignment component is 15% of the course's final grade and is only considered if the practice component has been passed. The chatbot was available to students for two weeks before the SQL final examination. Students did not receive specific instructions on how to work with the chatbot. Although the use of the chatbot was not mandatory, students were encouraged with a 10% increase of their grade in the assignment component if they used it and passed at least the first level.

A. Use Descriptive Scenario

The following is a detailed description of a sample user interaction. In Fig. 7, the conversation held by one of the students with the chatbot during the completion of one of the exercises available on the SQL learning platform is shown.

1. Chatbot proposes exercise: the statement of the exercise asks the student to find all the clients who live in Northampton.
2. Chatbot provides a hint: the chatbot suggests that the student type the necessary query to show the different tables in the database.
3. Student follows hint: the student types in the query box to visualise the different tables in the database.
4. Chatbot provides a hint: the chatbot suggests that the student type the necessary query to see the customer table.
5. Student follows hint: the student types in the query box to visualise the customer table.
6. Chatbot provides a hint: the chatbot tells the student that he/she has now all the information to solve the exercise.
7. The student proposes a solution: the student issues a query trying to solve the proposed exercise.
8. RDBMS provides feedback: the RDBMS evaluates the query and provides the console error code.
9. Chatbot provides feedback: then, the error is processed by the chatbot to display a more friendly message. In this case, the student has used a column that does not exist in the table queried in the database. As a result, the chatbot informs the student that the query is incorrect, indicating that the column to which the filter should be applied cannot be found.
10. Chatbot provides a hint: the previous hint is repeated.
11. Student proposes a solution: The student proposes another solution to the exercise.



Student Attempt Information ×

Queries

(3) SHOW tables

(5) DESCRIBE Customers

(7) SELECT * FROM Customers WHERE city = 'Northampton';

(11) SELECT * FROM Customers WHERE town = 'Northampton';

Errors

(8) SQLSTATE[42S22]: Column not found: 1054 Unknown column 'city' in 'where clause' (SELECT * FROM Customers WHERE city = 'Northampton';)

Conversation

(1) Hello! Let's do the following exercise: List all the customers who live in Northampton.

(2) Write the query needed to show the different tables in the database and identify the one you need.

(4) Now that you know the tables that compose the database, you need to know the structure of the one we need.

(6) Now that you know the fields in the table, use the ones you need to solve the query.

(9) The column you are trying to apply the filter to is not found.

(10) Now that you know the fields in the table, use the ones you need to solve the query.

(12) Correct! You are done with this exercise.

Fig. 7. Example of a student's conversation with the chatbot during an exercise (the numbers in parentheses have been included in the image editing to coincide with the numbering in the scenario description, but these numbers do not appear in the application).

12. Chatbot provides feedback: the RDBMS evaluates the query, and the instances of existing customers in the database who live in Northampton are listed. As a result, the chatbot informs the student that the query is correct, the exercise is considered solved, and the student proceeds to the next exercise.

B. Student Results in the Chatbot Environment

After monitoring the students' interactions with the chatbot, it can be seen that the SQL learning platform guides the students to solve the errors contained in their proposed solutions, allowing them to try again to define their solution on an unlimited number of opportunities. This approach can benefit students during the independent study of the Database subject since they cannot interact directly with the lecturer in such an easy and quick way. In Table I, some of the statistics automatically obtained from the application are collected. The first column corresponds to the exercise identifier. The second column shows the level of difficulty of the exercise: Beginner (A), Intermediate (B) and Advanced (C). The column 'Solved' shows the number of students who solved it successfully. The fourth column, 'Dropouts', shows the number of students who dropped out of the exercise without solving it. The column 'Errors' shows how many wrong queries a student makes on average before successfully solving the exercise. Finally, the column 'Queries' indicates how many queries on average the students use before finding the solution.

TABLE I. Statistics on Students' Attempts at Each Exercise

Ex.	Level	Solved	Dropouts	Error	Queries
E1	A	49	7	0,73	3,17
E2	A	54	2	0,34	2,37
E3	B	51	2	0,57	2,83
E4	B	51	2	0,09	1,53
E5	B	52	1	0,34	1,82
E6	B	42	11	1,09	5,00
E7	B	50	3	0,45	3,06
E8	C	36	15	1,52	4,21
E9	C	36	15	1,99	5,74
E10	C	47	4	0,49	2,73

For all the exercises, the number of students who managed to solve each one was much higher than those who dropped out. Similarly, the data helps identify which type of exercises are most difficult for the student. For example, exercise number 9 is the one with the least number of students who managed to solve (36), the one with the most dropouts (15), the one with the highest average number of errors (1.99) and the one that required the most attempts for those students who successfully solved it.

C. Impact on Students' Performance

To study the impact of chatbot use on students' performance, we considered student marks in the SQL final examination. 103 students took this examination, 54 of whom used the chatbot. Table II shows the correlation between the students' use or non-use of the chatbot and whether or not they passed the SQL final examination.

TABLE II. Use of the Chatbot Concerning Passing the Course

Chat box use	Passed the exam	Failed the exam	Total
Yes	23	31	54
No	9	40	49
Total	32	71	

We can see how of the 54 students who trained with the chatbot, 23 passed the SQL final examination (43%). Meanwhile, of the 49 who did not train with the chatbot, only 9 passed (18%). In other words, if a student trains with the chatbot, he/she is 25%more likely to pass the SQL final examination than if he/she does not train with the chatbot.

To verify whether the difference shown by the data is significant, we use the Chi-squared test and define the following null hypothesis:

- H_0: The fact that a student has trained with the chatbot is independent of passing the SQL final examination.

Therefore, the alternative hypothesis would be stated as follows:

- H_1: The fact that a student has trained with the chatbot is related to passing the SQL final examination.

To determine the dependency between passing the SQL final examination after training with the chatbot, we used the Chi-square test with a significance level of 0.05. As a result, we obtained a p-value of 8.2047 in the Chi-square test. This value is above the significance threshold of 0.05 ($X^2 > 3.8414$). Thus, we cannot accept the null hypothesis, and we assume that there is a relationship between training with the chatbot and passing the SQL final examination.

Below, we show the students' marks concerning whether they have used the chatbot or not. In Table III, it can be seen that the 5 students with the highest marks in the exam did use the chatbot. Similarly, the only ranges in which the students who did not use the chatbot outnumber those who did use it are in the three lowest mark ranges, especially in the last two.

TABLE III. Number of Students in Each Mark Range

Mark range	Chatbot use Yes	Chatbot use No
0.0-0.9	7 (13%)	14 (29%)
1.0-1.9	7 (13%)	11 (22%)
2.0-2.9	6 (11%)	7 (14%)
3.0-3.9	7 (13%)	5 (10%)
4.0-4.9	4 (7%)	3 (6%)
5.0-5.9	5 (9%)	3 (6%)
6.0-6.9	9 (17%)	4 (8%)
7.0-7.9	4 (7%)	2 (4%)
8.0-8.9	1 (2%)	0 (0%)
9.0-9.9	3 (7%)	0 (0%)
10.0	1 (2%)	0 (0%)

Although there is evidence that students who have used the chatbot have better marks than those who have not, it could be that the students who are more motivated and perform better in the subject are those who have used the chatbot.

To shed light on this issue, we will compare the marks of the students that worked with the chatbot with those of the previous year (Table IV). It can be seen that the marks, in general, have gone up one step regarding the previous year.

TABLE IV. Number and Percentage of Students in Each Mark Range Compared With Previous Year

Mark range	Chatbot students	Previous year students
0.0-0.9	7 (13%)	41 (30%)
1.0-1.9	7 (13%)	23 (17%)
2.0-2.9	6 (11%)	5 (4%)
3.0-3.9	7 (13%)	9 (7%)
4.0-4.9	4 (7%)	4 (3%)
5.0-5.9	5 (9%)	24 (18%)
6.0-6.9	9 (17%)	10 (7%)
7.0-7.9	4 (7%)	8 (6%)
8.0-8.9	1 (2%)	7 (5%)
9.0-9.9	3 (7%)	4 (3%)
10.0	1 (2%)	0 (0%)

While in the previous year, 47% of the students had the lowest marks (between 0.0 and 1.9), Among the students who used the chatbot in this year this indicator dropped to 26%. So, the students with the lowest marks are fewer.

In the three mark ranges between 2.0 and 4.9 we find a higher number of students in the chatbot group than in the previous year (31% vs. 14%). Although these are students who did not pass the exam, we see that the marks of failed students increased compared to the previous year.

Something similar happens with students who pass with the minimum mark (between 5 and 5.9). We can see that in the previous academic year, 18% of the students narrowly passed the exam. This percentage is somewhat lower in the chatbot course (9%), shifting part of this percentage to the next range (between 6 and 6.9) with 17% against 7% in the previous course.

Fig. 8 graphically shows this difference between the two cohorts. It can be seen how the red stripe (previous course) is higher in the 0.0-0.9 and 1.0-1.9 range column, while the blue stripe is higher in the 2.0-2.9, 3.0-3.9 and 4.0-4.9. Although these are still low marks, they do denote an improvement with respect to the students who obtain the lowest marks. In addition, the red stripe (previous course) is higher in the 5.0-5.9 range column, while the blue stripe (students who used the chatbot) is higher in the 6.0-6.9 than the red stripe, i. e. there is evidence that the tool has also helped to improve the level of students who passed the subject with the minimum mark. Finally, there is hardly any difference in the highest marks.

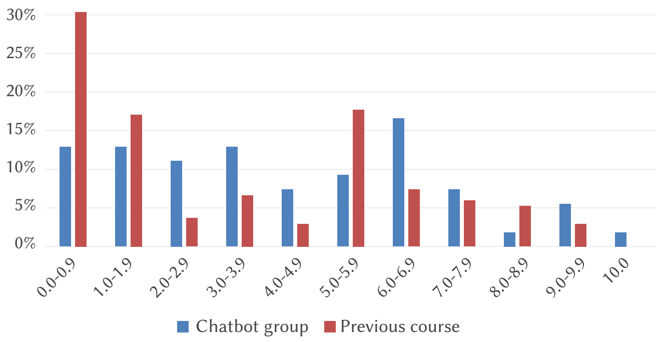


Fig. 8. Year comparison.

The analysis that follows from these results is the following:

- There is evidence that within the same academic year, students who practised with the chatbot get better marks than those who did not.
- Compared to the previous academic year, there is evidence that the improvements are more significant among students with lower and average marks.
- There is no evidence that students with medium-high marks move up in level and get the highest mark.
- This is consistent with comments from some students who, while enjoying the use of the chatbot, suggested the imple mentation of more complex queries.
- Therefore, there is evidence that the chatbot has helped students who were low performers in SQL language to improve their level and helped medium-high performers to brush up on their SQL skills.

D. User Acceptance Evaluation

A study to evaluate the user acceptance of the chatbot was conducted. Its design and its results are described below.

1. Survey Design

A survey to evaluate the user acceptance of the chatbot was designed with Google Forms to collect the users' opinions. The survey includes 15 questions to score several attributes according to the Technology Acceptance Model (TAM). The attributes analysed are Perceived Usefulness (PU), Perceived Ease-Of-Use (PEOU), Perceived Enjoyment (PE), Attitude (A) and Behavioural intention (BI). In addition, a question to assess the students' motivation for learning the theoretical concepts of the database course (LDBM) was included. These questions allow values in a Likert scale from 1–5, representing the user's dissatisfaction with value 1 and total compliance with value 5. Below are the questions included in the survey and their correspondence to each of the attributes mentioned above:

- Perceived ease-of-use (PEOU): Two questions were included to find out if the chatbot is easy to learn (Q1) and use (Q2).
- Perceived usefulness (PU): To measure this attribute, six questions have been included to find out if the chatbot enables users to learn the SQL syntax (Q3); assists in identifying the different errors that happen when running SQL queries (Q4); enables to learn how to issue simple SQL queries on a single table (Q5), queries that include SQL functions (e.g. min, max and others) (Q6), nested queries (Q7) and queries on multiple tables (Q8).
- Perceived enjoyment (PE): This attribute is measured through one question regarding whether the users had a positive experience using the chatbot (Q14).
- Attitude (A): Three questions have been included to find out if the users think that chatbots are useful learning platforms to help users solve problems, learn new things or be more productive (Q9), grasp computer engineering concepts from the academic curriculum (Q10), and learn how to manage databases (Q11).
- Behavioural intention (BI): For this attribute, two questions ask if the users would recommend using the chatbot to people interested in learning about databases (Q12) and if the users will use the chatbot to prepare themselves for further course assessments (Q13).
- Learning database motivation (LDBM): Lastly, students are asked whether if the chatbot encourages them to keep learning database concepts (Q15).

Five questions were also included to collect students' alias, age, gender, the latest exercise level they achieved to complete, and the number of calls they have been previously evaluated. These items enable us to classify user responses. Finally, their comments after using the chatbot are collected through a free text field. The survey is included as supplementary material to this manuscript.

2. Survey-Data Analysis

Firstly, 56 survey results were collected from the students. The results obtained and the analysis performed is available in a spreadsheet included as supplementary material. To check the validity of the data, an Alpha Cronbach test was conducted, obtaining very high confidence (0.98). The results obtained after the analysis of the measured attributes were significantly positive. The average values of the scores of the attributes are: PEOU 4.39, PU 3.98, PE 4.07, A 4.51, BI 4.36 and LDBM 4.1.

On the other hand, the answers collected have been classified according to the user's age and gender, the latest exercise level achieved and the number of calls the user has been evaluated. The results of the analysed data are included in Table V.

Although there are no significant differences, we can draw some interesting insights. First, the average data for the measured attributes are quite positive. Regarding gender, men score the PE attribute slightly lower (4) than women (4.57). Moreover, the analysis shows that the students who have previously taken one or more examinations score the PE attribute lower (3.69) than those who have taken no previous exam (4.19). The PE attribute also decreases to 3.75 for users above 20 years old, which may be related to the above insight. Furthermore, users who have not completed all the levels have scored lower on the PEOU (3.6), and PE (3.8) attributes. Finally, concerning the feedback provided by students, some ask for making clear several of the messages generated by the bot, others propose to increase the difficulty of the exercises, and others ask for allowing multi-line answer entries, among other improvements.

TABLE V. Results of the Survey With Students: the Measured Attributes (the Italic Font Shows the Chi-Squared Values)

User Profile	PEOU	PU	PE	A	BI	LDBM
Gender						
Man	4.26	4.28	*4*	4.45	4.43	4.10
Woman	4.43	4.43	4.57	4.86	4.86	4.71
Chi-squared	*0.63*	*0.63*	*0.59*	*0.65*	*0.65*	*0.61*
Number of examinations taken						
0	4.28	4.42	4.19	4.56	4.51	4.19
1 or more	4.31	*4*	*3.69*	4.31	4.38	4.15
Chi-squared	*0.63*	*0.65*	*0.76*	*0.66*	*0.66*	*0.62*
Age						
<=20	4.36	4.42	4.25	4.64	4.58	4.19
>20	4.15	4.15	*3.75*	4.25	4.30	4.15
Chi-squared	*0.64*	*0.65*	*0.76*	*0.67*	*0.67*	*0.62*
Exercise levels completed						
A & B levels	*3.60*	4.40	*3.80*	4.60	5	***4***
All levels	4.35	4.31	4.10	4.49	4.43	4.20
Chi-squared	*0.54*	*0.65*	*0.57*	*0.67*	*0.71*	*0.59*
Students						
All	4.39	3.98	4.07	4.51	4.36	4.1

V. Discussion

This research is aimed at supporting students in learning and training with SQL. From the results collected, we can report positive evidence using the chatbot for both students and lecturers.

Previous work in the literature related to using chatbots and learning a programming language focuses on analysing chatbot usability and acceptance. However, it does not analyse the impact of chatbot use on learner outcomes [30], [33], [34]. In this paper, we could analyse the impact of the chatbot on student performance. We observed that a student who passes the exam is much more likely to have trained with the chatbot than one who has not. In other words, the students who use the chatbot obtain better results than those who did not use it. Of course, that does not necessarily mean that training with the chatbot is the only variable involved in this increase in passing. However, there are likely other characteristics of the students that have pushed them to use the chatbot. Nevertheless, we do consider these numbers to be positive evidence of the benefit of the chatbot.

The positive impact of the chatbot can also be confirmed with the students' evaluation carried out through the TAM. All the TAM attributes analysed were in the upper range (between 4 and 5), except for perceived usefulness at 3.98. It is worth noting that students also had a high motivation to learn databases (4.1 out of 5). Student motivation is fundamental to their academic success [41]. That could also point to the significant difference found in passing the final SQL examination between the students who used the chatbot and those who did not. The students who used the chatbot are the most motivated to learn databases and therefore used the chatbot. Besides, some of their survey responses indicated that the chatbot could be enriched with more complex queries, suggesting that they were eager to further work with the chatbot. This conclusion is in line with the findings of Munday [34], who reported that, after finishing the course, 10% of her students continued to use the chatbot to improve their language skills. Future work along these lines could aim to engage less motivated students.

From the lecturers' point of view, the learning platform assists in monitoring students' progress. The analysis of learning logs can be very relevant for lecturers when assessing computer programming tasks [42]. From the records of the students' attempts at the different exercises, the lecturer involved in the study was able to draw beneficial conclusions:

- The contents of the course that students mastered, based on the queries they solved successfully and with fewer attempts.
- The learning aspects that must be reinforced, according to the queries that were more difficult for the students and had a lower success rate.
- It allows us to know which errors were repeated more frequently when trying to solve the exercises.
- It has allowed us to know the results of the advice provided by the chatbot that has been more or less useful for the students to solve the problems.

Therefore, we encourage lecturers to implement learning platforms based on chatbots to support students learning program-ming languages.

VI. Conclusions

Learning SQL is one of the critical activities for any computer engineer, but it often presents difficulties for students. For this reason, this paper presents a chatbot embedded in a learning platform to help students in databases to perform SQL queries correctly. The chatbot was implemented and made available to the students of a university course on Databases in the second year of the degree in Computer Engineering. The students use it for two weeks before the final SQL examination. The results, so far, are promising because:

- The chatbot helps students learn SQL when the lecturer is not present by adapting to the needs of individual students based on their errors.
- Our data analysis revealed a high acceptance of the chatbot by the students.
- There is evidence that the chatbot helped students to achieve better results, specially to students with both average and low marks.
- The SQL learning platform provides detailed information to lecturers about their students' learning process that they can use to improve it.

As future work, we plan to conduct further studies with more students to collect more data that will enable us to apply Machine Learning (ML) algorithms to improve various aspects of the SQL learning platform. Thanks to the automatic ML algorithms that IBM Watson incorporates and the lecturers' adjustments, the precision of the answers and advice offered by the chatbot will be progressively improved. Also, implement higher levels of difficulty that also motivate students with higher marks.

Appendix

A. Survey to Evaluate Chatbot Acceptance

Students must answer all questions by assigning a value according to the following 5-point Likert scale: (1) Strongly Disagree (2) Disagree (3) Neither Agree nor Agree (4) Agree (5) Strongly Agree

1. Perceived Ease-Of-Use (PEOU)

- I consider the SQL chatbot to be easy to learn (Q1)
- I consider the SQL chatbot to be easy to use (Q2)

2. Perceived Usefulness (PU)

- I consider that the SQL chatbot allows me to learn the syntax of the SQL language (Q3)
- I consider that the SQL chatbot helps me to distinguish the different types of errors that occur when launching queries (Q4)
- I consider that the SQL chatbot allows me to learn how to launch simple SQL queries on a single table (Q5)
- I consider that the SQL chatbot allows me to learn how to launch SQL queries that include functions (e.g. min, max, etc.)(Q6)
- I consider that the SQL chatbot allows me to learn how to launch nested SQL queries (Q7)
- I consider that the SQL chatbot allows me to learn how to launch SQL queries involving several tables with joins, etc (Q8)

3. Attitude (A)

- I consider chatbots to be useful tools that can help us to solve problems, learn new things or be more productive (Q9)
- I consider it interesting to use chatbots for the acquisition of Computer Engineering concepts that are studied at the University (Q10)
- I like the idea of using chatbots to learn how to manage databases (Q11)

4. Behavioural Intention (BI)

- I would recommend this tool to anyone interested in learning how to use databases (Q12)
- I will use this application to prepare myself for the Database subject assessments (Q13)

5. Perceived Enjoyment (PE)

- When I use the SQL chatbot I have a good time (Q14)

6. Learning Database Motivation (LDBM)

- Using the SQL chatbot makes me more interested in learning databases (Q15)

Acknowledgement

This research was funded by Spanish National Research Agency (AEI), through the project CRÊPES (ref. PID2020-115844RB-I00) with ERDF funds.

The authors would like to thank Daniel Mejías, who con-tributed to the development of the chatbot, and the University of Cádiz for its support under its Innovation in Education pro-gramme (ref. sol-202000162042-tra), and through the University Research Institute for Sustainable Social Development (INDESS).

References

[1] J. Cárdenas-Cobo, A. Puris, P. Novoa-Hernández, Á. Parra-Jiménez, J. Moreno-León, D. Benavides, “Using scratch to improve learning programming in college students: A positive experience from a non-weird country,” *Electronics*, vol. 10, no. 10, p. 1180, 2021.

[2] F. J. García-Peñalvo, V. Abella-García, A. Corell, M. Grande, “La evaluación online en la educación superior en tiempos de la covid-19,” *Education in the Knowledge Society*, vol. 21, no. 12, 2020.

[3] A. Balderas, J. A. Caballero-Hernández, “Analysis of Learning Records to Detect Student Cheating on Online Exams: Case Study during COVID-19 Pandemic,” in *Proceedings of the Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM’20)*, 2020, ACM.

[4] S. Tamayo, D. Pérez-Marín, “¿Qué esperan los docentes de los agentes conversacionales pedagógicos?,” *Teoría de la Educación. Educación y Cultura en la Sociedad de la Información*, vol. 18, no. 3, pp. 59–85, 2017.

[5] G. D’Aniello, A. Gaeta, M. Gaeta, S. Tomasiello, “Self-regulated learning with approximate reasoning and situation awareness,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, no. 1, pp. 151–164, 2018.

[6] B. A. Shawar, E. Atwell, “Using dialogue corpora to train a chatbot,” in *Proceedings of the Corpus Linguistics 2003 conference*, 2003, pp. 681–690.

[7] E. K. Morales-Urrutia, J. M. Ocaña, D. Pérez-Marín, “How to integrate emotions in dialogues with pedagogic conversational agents to teach programming to children,” in *Innovative Perspectives on Interactive Communication Systems and Technologies*, IGI Global, 2020, pp. 66–91.

[8] J. Yin, T.-T. Goh, B. Yang, Y. Xiaobin, “Conversation technology with micro-learning: the impact of chatbot-based learning on students’ learning motivation and performance,” *Journal of Educational Computing Research*, vol. 59, no. 1, pp. 154–177, 2021.

[9] A. Beaulieu, *Learning SQL: master SQL fundamentals.* ” O’Reilly Media, Inc.”, 2009.

[10] J. V. Murillo, S. B. Chavarría, S. M. Rivera, “Herramienta asistida por computadora para la enseñanza del álgebra relacional en bases de datos,” *Uniciencia*, vol. 26, no. 1-2, pp. 179–195, 2012.

[11] F. Carrillo Chaves, “Informe asignaturas de bases de datos en estudios de grado en ingeniería informática en universidades españolas, 2018.” http://hdl.handle.net/10498/ 20657, 2018.

[12] C. R. Jaimez-González, A. Palma-Hernández, “An interactive online training course for sql beginnersbeginners,” *International Journal on Integrating Technology in Education (IJITE)*, vol. 6, no. 4, pp. 1–9, 2017.

[13] A. Cumming, *sqlcourse.com Interactive Online SQL Training*, 2009 (accessed July 8, 2020).

[14] R. Baggetun, B. Wasson, “Self-regulated learning and open writing,” *European Journal of Education*, vol. 41, no. 3-4, pp. 453–472, 2006.

[15] P. Brusilovsky, S. Sosnovsky, M. V. Yudelson, D. H. Lee, V. Zadorozhny, X. Zhou, “Learning sql programming with interactive tools: From integration to personalization,” *ACM Transactions on Computing Education (TOCE)*, vol. 9, no. 4, pp. 1–15, 2010.

[16] H. Bansal, R. Khan, “A review paper on human computer interaction,” *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 8, no. 4, pp. 53–56, 2018.

[17] G. Garcia Brustenga, M. Fuertes Alpiste, N. Molas Castells, *Briefing paper: chatbots in education.* Barcelona: eLearn Center. Universitat Oberta de Catalunya (UOC), 2018.

[18] S.-W. Hsieh, “Effects of cognitive styles on an msn virtual learning companion system as an adjunct to classroom instructions,” *Educational Technology & Society*, vol. 14, no. 2, pp. 161–174, 2011.

[19] J. Weizenbaum, “Eliza—a computer program for the study of natural language communication between man and machine,” *Communications of the ACM*, vol. 9, no. 1, p. 36–45, 1966.

[20] P. B. Brandtzaeg, A. Følstad, “Why people use chatbots,” in *Internet Science*, Cham, 2017, pp. 377–392, Springer International Publishing.

[21] V. Di Lecce, M. Calabrese, D. Soldo, A. Giove, “Semantic management systems for the material support of e-learning platforms,” *Journal of e-Learning and Knowledge Society*, vol. 6, no. 3, pp. 61–70, 2010.

[22] Apple Computer, “Siri.” https://www.apple.com/siri, 2010. Last accessed 1 April 2022.

[23] IBM, “Ibm watson.” https://www.ibm.com/cloud/ watson-assistant, 2011. Last accessed 1 April 2022.

[24] Google, “Google assistant.” https://assistant.google. com/, 2016. Last accessed 1 April 2022.

[25] X. Luo, S. Tong, Z. Fang, Z. Qu, “Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases,” *Marketing Science*, vol. 38, no. 6, pp. 937–947, 2019.

[26] A. De Angeli, G. I. Johnson, L. Coventry, “The unfriendly user: exploring social reactions to chatterbots,” in *Proceedings of the international conference on affective human factors design, london*, 2001, pp. 467–474, Citeseer.

[27] Q. Zhi, R. Metoyer, “Gamebot: a visualization-augmented chatbot for sports game,” in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–7.

[28] D. Calvaresi, J.-P. Calbimonte, E. Siboni, S. Eggenschwiler, G. Manzo, R. Hilfiker, M. Schumacher, "Erebots: privacy-compliant agent-based platform for multi-scenario personalized health-assistant chatbots," *Electronics*, vol. 10, no. 6, p. 666, 2021.

[29] W. Ham, N. Plainstow, "New deal for communities."splodge" new deal for communities chatbot assistant," 2005.

[30] F. Colace, M. De Santo, M. Lombardi, F. Pascale, A. Pietrosanto, S. Lemma, "Chatbot for e-learning: A case of study," *International Journal of Mechanical Engineering and Robotics Research*, vol. 7, no. 5, pp. 528–533, 2018.

[31] R. Scheepers, M. Lacity, L. Willcocks, "Cognitive automation as part of deakin university's digital strategy," *MIS Quarterly Executive*, vol. 17, no. 2, pp. 89–107, 2018.

[32] A. K. Goel, L. Polepeddi, "Jill watson: A virtual teaching assistant for online education," in *Learning Engineering for Online Educationn*, C. Dede, J. Richards, B. Saxberg Eds., Routledge: Taylor & Francis, 2016, ch. 7, pp. 120–143.

[33] P. Yeves-Martínez, D. Pérez-Marín, "Prof. watson: A pedagogic conversational agent to teach programming in primary education," in *Multidisciplinary Digital Publishing Institute Proceedings*, vol. 31, 2019, p. 84.

[34] P. Munday, "The case for using duolingo as part of the language classroom experience," *RIED: revista iberoamericana de educación a distancia*, vol. 19, no. 1, pp. 83–101, 2016.

[35] S. Brown, *The art of visualising software architecture: communicating software architecture with sketches, diagrams and the C4 model.* Lean Publishing, 2016.

[36] M. Stauffer, *Laravel: Up & Running: A Framework for Building Modern PHP Apps.* O'Reilly Media, 2019.

[37] F. Bracci, A. Corradi, L. Foschini, "Database security management for healthcare saas in the amazon aws cloud," in *2012 IEEE Symposium on Computers and Communications (ISCC)*, 2012, pp. 812–819, IEEE.

[38] S. Tilkov, S. Vinoski, "Node. js: Using javascript to build high-performance network programs," *IEEE Internet Computing*, vol. 14, no. 6, pp. 80–83, 2010.

[39] C. Müller, M. Stahl, M. Alder, M. Müller, "Learning effectiveness and students' perceptions in a flexible learning course," *European Journal of Open, Distance and E-learning*, vol. 21, no. 2, pp. 44–52, 2018.

[40] E. Kim, L. Rothrock, A. Freivalds, "The impact of gamification on the motivation and performance of engineering students through the lens of self-determination theory," *International Journal of Engineering Education*, vol. 36, no. 3, pp. 1117–1131, 2020.

[41] H. Afzal, I. Ali, M. Aslam Khan, K. Hamid, "A study of university students' motivation and its relationship with their academic performance," *International Journal of Business and Management*, vol. 5, no. 4, pp. 80–88, 2010.

[42] X. Fu, A. Shimada, H. Ogata, Y. Taniguchi, D. Suehiro, "Real-time learning analytics for c programming language courses," in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 2017, pp. 280–288.

Antonio Balderas

Antonio Balderas received his MSc degree in computer science and his PhD degree from the University of Cadiz, Spain. He is currently with the University of Cadiz, and works as an Assistant Professor in the Department of Computer Engineering and as a Researcher with the Software Process Improvement and Formal Methods Group. He was a project manager in different Spanish IT companies. His research interests include technology-enhanced learning and creative computing.

Rubén Baena-Pérez

Rubén Baena-Pérez received a degree in Fine Arts from the University of Seville and is a PhD student in Computer Science at the University of Cadiz. He is currently a professor at the University of Cadiz, Spain, and a researcher in the Software Process Improvement and Formal Methods Group. His research fields are technology-enhanced learning and creative computing.

Tatiana Person

Tatiana Person has a Master's Degree in Software Engineering from the University of Cadiz (UCA). She has worked as an as-sistant lecturer in the Department of Computer Engineering of the UCA for 4 years and she is a Researcher in the Software Process Improvement and Formal Methods Group. Currently, she is Computer Engineer with Tecnobit (Oesía Group) Engineering Company and she is doing her PhD on how to bring data analytics techniques and mobile software development closer to noncoders.

José Miguel Mota

José Miguel Mota received the master's degree in computer science with the Universitat Oberta de Catalunya and a PhD degree from the University of Cadiz. He is currently with the University of Cadiz, and works as an Associate Lecturer in the Department of Computer Engineering and as a Researcher with the Software Process Improvement and Formal Methods Group. His current research interests include technology-enhanced learning, augmented reality, virtual reality and learning analytics. He has published several papers and book chapters in these fields.

Iván Ruiz-Rube

Iván Ruiz-Rube received his MSc degree in software engineering from the University of Seville and a PhD degree in computer science from the University of Cádiz. He has been a Development Engineer with Everis and Sadiel ICT consulting companies. He is currently an associate professor with the University of Cádiz, Spain. His fields of research are software process improvement and technology-enhanced learning. He has co-authored 26 papers both in JCR and SJR indexed journals, as well as 3 book chapters.

Supporting Skill Assessment in Learning Experiences Based on Serious Games Through Process Mining Techniques

Juan Antonio Caballero-Hernández[1]*, Manuel Palomo-Duarte[2], Juan Manuel Dodero[3], Dragan Gaševic[4]

[1] EVALfor research group, University of Cadiz, Puerto Real (Spain)
[2] Department of Computer Science, University of Cadiz, Puerto Real (Spain)
[3] Department of Computer Science, University of Cadiz, Puerto Real (Spain)
[4] Faculty of Information Technology, Monash University, Melbourne (Australia)

* Corresponding author. juanantonio.caballero@uca.es

Received 26 September 2021 | Accepted 11 February 2023 | Early Access 9 May 2023

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

Learning experiences based on serious games are employed in multiple contexts. Players carry out multiple interactions during the gameplay to solve the different challenges faced. Those interactions can be registered in logs as large data sets providing the assessment process with objective information about the skills employed. Most assessment methods in learning experiences based on serious games rely on manual approaches, which do not scalewell when the amount of data increases. We propose an automated method to analyse students' interactions and assess their skills in learning experiences based on serious games. The method takes into account not only the final model obtained by the student, but also the process followed to obtain it, extracted from game logs. The assessment method groups students according to their in-game errors and in-game outcomes. Then, the models for the most and the least successful students are discovered using process mining techniques. Similarities in their behaviour are analysed through conformance checking techniques to compare all the students with the most successful ones. Finally, the similarities found are quantified to build a classification of the students' assessments. We have employed this method with Computer Science students playing a serious game to solve design problems in a course on databases. The findings show that process mining techniques can palliate the limitations of skill assessment methods in game-based learning experiences.

Keywords

Educational Process Mining, Game-Based Learning, Learning Analytics, Model Discovery, Serious Games, Skill Assessment.

DOI: 10.9781/ijimai.2023.05.002

I. Introduction

Serious games are considered to be those games which have purposes beyond entertainment [1]. The employment of serious games in educational contexts is promising to create and develop learning processes where students are actively involved. A lot of research on the positive impact and outcomes associated with playing serious games can be found in the literature [2]. Although serious games are widely employed in online learning, the methods for their assessment still rely on manual approaches, which are scarce in details of the assessment of the learning outcomes, have scalability problems, and lack automated and semi-automated support [3]. During online gameplay, players can carry out diverse interactions according to the features of the game, e.g.: movements, such as jumping or running, selections of a text option in a conversation, selection of an item using the pointer, etc. These interactions can remain in storage records, databases, or log files, thus resulting in data sets that include objective information about the skills employed during the game. The analysis of such large data sets can lead to scalability problems when using manual methods of assessment.

The processes of Learning Analytics (LA) can palliate these limitations through data-driven analysis. LA is "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [4]. The interactive nature of serious games makes them a significant source of data, tracking user interactions and storing them as sequential events in a log. Among the sequence analysis techniques, process mining can be used to discover, monitor and improve the actual processes by extracting knowledge from an event log [5]. As a discipline, process mining is situated between computational intelligence, data mining, and process modeling and analysis. Due to the sequential nature of

Please cite this article as: J. A. Caballero-Hernández, M. Palomo-Duarte, J. M. Dodero, D. Gaševic, "Supporting Skill Assessment in Learning Experiences Based on Serious Games Through Process Mining Techniques", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 6, pp. 146-159, 2024, http://dx.doi.org/10.9781/ijimai.2023.05.002

game interactions and the identified limitations of their assessment, the techniques for their event-based data analysis are considered to provide the basis of an automated method to support skill assessment in learning experiences based on serious games.

This paper proposes a method based on a combination of process mining techniques to analyse the logs produced by a serious game. We have developed process models and carry out a conformance check to validate them with particular event logs. Our method conducts a performance comparison using assessment models or profiles, inspired by existing assessment integration approaches [6]. Assessment models can be considered as good examples of the behaviour that can be imitated by the students. The proposed method provides an assessment model as well as indicators to measure the gap between the behaviour represented by the assessment model and the behaviour of the students. The method conducts a more detailed assessment through behaviour analysis to identify similarities and differences between each student and those students who are most successful in their learning outcomes. The method considers objective evidence from the process carried out by the student during the experience, extracted from the game logs.

In order to validate our proposal, a case study was carried out in the context of a Computer Science degree program. More than 100 students enrolled in a course on databases played a serious game and designed a conceptual data specification through an Entity–Relationship (E/R) diagram. In playing the game, the students were expected to use a specific skill for the analysis and design of relational databases, focused on the learning outcome related to the E/R diagram modelling, namely "knowledge to produce a logical and conceptual design of a database."

The rest of this paper is organized as follows. In Section II we place the subject of serious games in its context, as well as learning assessment methods, as found in the literature. Section III introduces process mining and defines its main concepts. Then, the research questions and the proposal are described in Section IV. The conducted case study is detailed in Section V. A discussion of the results of implementing our process mining proposal is presented in Section VI. Finally, we list the conclusions of the study and identify future lines of work in Section VII.

II. Learning Assessment in Serious Games

Assessment in the educational process validates the acquisition of skills by the students. Allen defines assessment as "the use of empirical data on student learning to refine programs and improve student learning" [7]. Despite the variety of assessment methods found in the literature, most of their implementations rely on manual assessment [3]. Manual assessment covers all those processes supported by traditional approaches to assessing student skills, such as instructor observations, which might be based on subjective assumptions, and traditional tests, where the answers to each question are seen as independent data points. However, learning and succeeding in a complex and dynamic world is not easily measured by multiple-choice responses on a simple knowledge test [8]. Manual assessment can suffer from problems in assessing large data sets, due to scalability limitations and the lack of automated and semi-automated mechanisms to support the assessment. In addition, serious games are commonly employed with formative aims, while the assessment of the acquired skills is implemented through external tools with predefined answers, resulting in possible omissions of relevant information.

Some well-known areas related to LA are usually employed to assessment in serious games to palliate some of these limitations. Evidence-centred assessment design (ECD) is a framework to provide language, concepts and knowledge representations for designing and delivering learning assessments, organized around the evidentiary argument an assessment is meant to embody [9]. ECD contains a conceptual assessment framework (CAF) layer, widely used for educational assessment development and considered the blueprint for an assessment [10]. CAF is divided into models, where each model provides the answer for critical questions related to the assessment process. Student models define one or more variables related to the skills we wish to measure. Evidence models provide concrete instructions for analysing and measuring the variables defined within the student model. Task models describe the situations in which to obtain the evidence needed for the evidence models.

Following the ECD approach, stealth assessment is an embedded and in-process assessment, usually focused on formative aims. Stealth assessment aims to support learning and keep the student engaged in the activity while removing or reducing test anxiety without sacrificing reliability and validity [8]. Stealth assessment is seamlessly included in the educational process. It represents a quiet process by which student's interactions involving the levels of the relevant skills are stored in a dynamic model [11].

One widely used tool for stealth assessment is a Bayesian network [8]. Bayesian networks can be used within student models to handle uncertainty by using probabilistic inference to update and improve belief values (i.e., regarding student skills). Bayesian networks have been used in assessment systems where player interactions are captured during the game and related key indicators provide evidence for the skills employed [12]. Bayesian networks have also been used to assess students' performance in intelligent tutoring systems. Three constraint-based intelligent tutoring systems focused on database education are presented by Mitrovic et al. [13]. That approach provides feedback to students according to a description of the basic principles and concepts in the domain. One of these intelligent tutoring systems, called KERMIT, focuses on database modeling. KERMIT has been evaluated to prove its effectiveness according to the students' results [14].

An important and broadly used technique to model processes is Bayesian Knowledge Tracing (BKT) [15]. BKT is based on hidden Markov models and has been extensively used to perform assessments in intelligent tutoring systems [16]. BKT assumes that the student skills are represented as a set of binary variables. Each variable represents a skill that can be mastered by the student or not.

On the one hand, data-centred approaches tend to be agnostic as to the process: data mining, statistics, and machine learning do not consider end-to-end process models. On the other hand, process science approaches are process-centric but usually focused on modeling instead of discovering knowledge from the event data. Process mining is a mixed approach, between model-based process analysis and data-centred analysis [17]. Process mining seeks to confront the event data (i.e. the gathered evidence) with process models (generated automatically or hand made). In addition, process mining provides a more comprehensive and detailed picture of the structure of events that occur during a learning process, instead of having to aggregate process data into the frequencies or probabilities of events [18]. We consider that the unique position of process mining makes it a powerful tool for exploiting the growing availability of data from serious games and analyse in detail the processes performed by the students to be considered in the assessment.

III. Process Mining in Educational Environments

Process mining is considered a link between data mining and business process modeling and analysis [5]. An event log is the initial input in process mining. In general, an event log can be seen as a collection of cases and a case can be seen as a trace of events. Any event must include some mandatory fields: a unique identificator per process instance called "CaseID", the corresponding "activity" and its "timestamp".

There are three types of process mining techniques: discovery, conformance and enhancement [17]. Discovery techniques use an event log produced by any type of process as input to produce a model without using any a priori information. Conformance checking techniques compare an existing model to an event log of the same process. Enhancement techniques are focused on improving or extending a process model through the information stored in an event log of the same process.

There are four desired criteria for the quality of a discovered model: fitness, precision, generalization and simplicity. Fitness measures how well the model is able to replay most of the traces in a log. Precision measures the model's acceptance of unrelated behaviour, so that a model with low precision is underfitting, thus enabling a completely different behaviour from the event log. Generalization measures how well the model can generalize the analysed behaviour: a model with a low generalization is too specific and too adapted to the behaviour of the event log. Lastly, simplicity uses Occam's Razor: "one should not increase, beyond what is necessary, the number of entities required to explain anything" [19]. Therefore, the discovered model should be the simplest model that represents the event log.

In those cases where the models are rather extreme, the scores for the quality criteria will be evident. However, it is more difficult to judge the quality of a model in realistic contexts. Conformance checking enables finding similarities and/or discrepancies between the modeled behaviour (discovered model) and the observed behaviour (event log) [17]. Conformance checking relates the events in the event log to the activities in the process model and compares them. The comparison is carried out by a "replay": a process to check whether each log trace can be simulated through the states of the discovered model. Using a replay, the fitness can be quantified to measure the similarities between the model and the event log.

Model discovery and conformance checking techniques enable the analysis of processes to provide behaviour models and compare their performance. However, correlating the different characteristics of a process can be essential to conducting a more refined analysis. These characteristics can be based on different perspectives, such as the control flow (i.e. the next interaction), the data flow (i.e. the age of the player), the time (i.e. the duration of the game) or the resource (i.e. the player who performed the action) [20]. In the assessment of students during a learning experience based on a serious game, it can be important to detect the errors made in the game and how they determined the final result of the game. Therefore, a correlation between the in-game errors and the results could determine the most decisive in-game errors.

Process mining has been widely used in multiple domains, including educational environments. Actually, an accepted term to refer to the use of process mining in educational environments is Educational Process Mining (EPM) [21]. EPM is focused on the use of the event logs registered by educational environments to discover, analyse and detect the most common behaviours performed during the learning process.

Different applications of EPM in higher education can be found in the literature. Model discovery has been used in multiple learning experiences, such as relating the students' performance to their studying behaviour or assessing wiki contributions during a collaborative experience [22], [23]. Conformance checking has also been used in studies focused on EPM, such as the introduction of the event-centred view of a process as a generally applicable approach for providing closer links between qualitative and quantitative research methods [24]. The work of Bannert and Reimann stands out in its use of EPM to identify process patterns in self-regulated learning [25],[26]. First, model discovery is employed to model the behaviour of different student profiles. A student profile can be considered as a group of students who share similar characteristics, such as outcomes or in-game errors. Comparing the students' profiles makes it possible to detect differences in their behaviour. Finally, a theoretical expert model is compared against the empirical trace data, using conformance checking.

IV. Process Mining Support of Skill Assessment in Serious Games

In this section, we present a method based on process mining techniques to support assessment in a learning experience based on serious games. First, the research question and sub-questions of this study will be presented. Then, the process followed to use the proposed process mining techniques and answer the research question and sub-questions will be given in detail.

A. Research Questions

Some limitations have been identified in the manual methods of assessing serious games: scalability problems, assessments that lack details, and a lack of automatic and semi-automatic mechanisms for the assessments. Considering these limitations and the possibilities provided by process mining, the main research question of our study is: *can process mining techniques support a scalable assessment in learning experiences based on serious games?* This question has been divided into the following sub-questions.

- RQ1: Can process mining techniques identify the most decisive in-game errors for specific student profiles?
- RQ2: Can process mining techniques detect similarities and differences between the performance of specific student profiles?
- RQ3: Can process mining techniques allow students to be assessed according to their performance during the game?

B. The Proposed Assessment Method

The proposed method addresses these research sub-questions by using different process mining techniques through the ProM open source framework, because it provides multiple functionalities for analysing large data sets and supports all the techniques included in our proposal [27]. The newest version (6.9, at the time this experiment was conducted) of ProM was used.

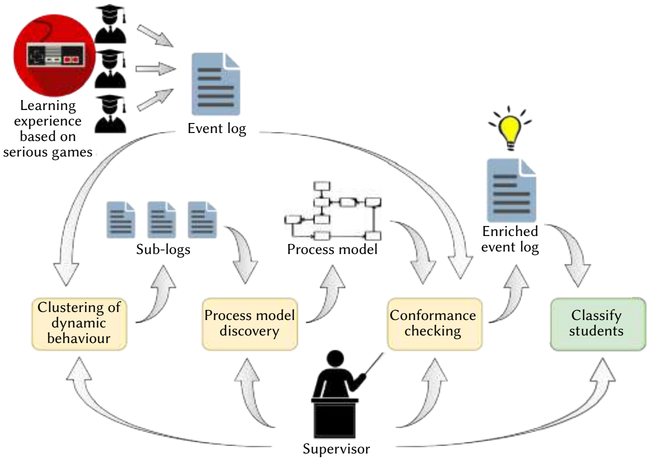


Fig. 1. Method to support the assessment of a learning experience based on serious games applying process mining techniques.

A case study has been conducted, using an event log produced by a serious game as the input for the proposed method, illustrated in Fig. 1. First, decision trees are generated to obtain trace clusters. This clustering makes it possible to detect different student profiles through a correlation analysis between students' outcomes and several types of in-game errors. Other existing approaches use different clustering techniques, such as agglomerative hierarchical clustering or k-means [28]. However, applying these techniques for trace clustering could not provide a clear insight into the characteristics of the process that are common to all the traces within the same cluster. Decision trees clearly highlight the discriminative characteristics, providing the reason why a certain log trace belongs to a given cluster and not to another [20].

Then, two subsets of the event log are obtained based on the previous clusters. These sub-logs contain detailed in-game errors for specific student profiles: the most successful students (MSS) and the least successful students (LSS). Behaviour similarities and differences between these student profiles are detected through model discovery techniques. The MSS model is used as an assessment model to be compared with each student's behaviour during gameplay through conformance checking. This comparison provides quantified data that represent the level of behaviour similarity between each student and the assessment model. This quantified data was added to the event log to enrich it. Finally, the enriched event log is explored to classify students according to their behaviour similarity in comparison to the assessment model. The stages of the proposed method are given in detail in the next subsections.

C. Clustering of Dynamic Behaviour

First, we aim to identify the most decisive in-game errors for the outcomes obtained by specific student profiles. This process is based on a successful system for clustering dynamic behaviour proposed and implemented in ProM [20]. Our starting point is an event log including all student gameplays in a learning experience based on serious games. During the game, the interactions performed by the students are stored in a log where each instance is a trace of events. These events must include some mandatory fields: the corresponding CaseID (a unique identificator per game experience of a student), the activity (performed interaction) and its timestamp. Depending on the context, additional fields can be included (i.e. grades achieved by a specific student, academic year, etc.). The in-game errors and final outcomes of each student were used during this case study, which is described in detail in Section V.

Starting from the event log, a process analysis can determine a so-called "analysis of a use case", which is defined as a triple $A = (c_r, C_d, F)$ consisting of [20]:

- Let C be the universe of characteristics,
    - c_r is a dependent characteristic defined as $c_r \in C \backslash C_d$,
    - C_d is a set of independent characteristics defined as $C_d \subseteq C \backslash \{c_r\}$,
- Let ε be the universe of events,
    - F is an event-selection filter defined as $F \subseteq \varepsilon$, which selects the events that are retained for the analysis.

Then, an analysis of a use case has to be defined to select a dependent characteristic, independent characteristics and an optional event filter. The behaviour analysis results in a decision tree whose purpose is to relate the dependent characteristic to the independent characteristics [20]. The ProM implementation of this system constructs decision trees relying on the algorithm C4.5 developed in the Weka toolkit [28].

Finally, the obtained decision tree can be used to cluster the executions of process instances with similar outcomes. In this tree, each trace is linked to one single instance that is associated with one leaf. All log traces associated with the same leaf are grouped within the same cluster, producing the same number of clusters as there are leaves in the tree. Event logs that include traces from multiple students are used to present a range of behavioural variability. Splitting the event log and grouping similar traces enables discovering partial models that are easier to understand and more representative than a discovered model produced by the whole event log. The employed implementation provides two well-known discretization techniques: equal-width binning and equal-frequency binning [29].

This process is structured as follows:

```
1.  Input: Event log
2.  Define an analysis of a use case ⟶ Use case
3.  Analysis technique ⟶ Decision tree
4.  Clustering decision tree ⟶ Sub-logs
5.  if Sub-logs include [Generic behaviour]
6.      Refine event log
7.      Go to step 2
8.  else //Sub-logs include [Specific behaviour]
9.      Sub-logs include [Student profiles]
```

D. Discovery of a Process Model

We aim to detect the similarities and differences between the behaviours of specific student profiles. Based on the results of the clustering of the dynamic behaviour, clusters are selected to create sub-logs that represent these student profiles. Process patterns of the student profiles are analysed through discovery techniques: Inductive Miner (IM) and its variant – infrequent (IMi) [30].

First, IM aims to discover from any given event log a set of process models that fit the observed behaviour. Then, IMi adds infrequent behaviour filters to all steps of IM [5]. If the model obtained with IM is not precise enough at evaluating the quality criteria (overfitting, underfitting, etc.), the miner is applied again using the same sub-log selecting the IMi. This process is iterated until precise models for all the student profiles are discovered. Finally, the models can be compared through visual inspection. The described process is structured as follows.

```
1.  Input: Sub-log for a student profile
2.  Mine using IM ⟶ Discovered model
3.  if Discovered model is [Imprecise]
4.      Mine using IMi ⟶ Discovered model
5.      Go to step 3
6.  else //Discovered model is [Precise]
7.      if [Pending student profiles]
8.          Go to step 1
9.      else //Not [Pending student profiles]
10.         Compare models
```

E. Conformance Checking

Finally, we aim to support the assessment process according to the performance during the game. Previously discovered models for specific student profiles are the input, along with an event log, for the conformance checking techniques. These techniques make it possible to compare the behaviour of a process model and the behaviour recorded in an event log. Replay is the process to quantify this comparison. It simulates the event log cases given a discovered process model, observing each log trace and showing the logic between the activities in the model.

Our method integrates the assessment process conducting a comparison of performance using assessment models or profiles, an approach proposed by Hainey et al. [6]. Using the discovered model concerning the MSS as the assessment model, the comparison is conducted replaying all the traces of the event log to the MSS model. As a result, the alignment and a fitness value to quantify how each trace (student) fits into the model (most successful students) are obtained.

Fitness is the most suitable criterion for conformance checking techniques as it measures how well the model is able to replay most of the traces in a log. Replay techniques can quantify the fitness, finding an alignment of traces in the event log with the control flow of the process models. The nodes of a model can hold one or more tokens and a set of directed arrows that represent the transition between the nodes. Transitions are enabled as soon as all the nodes connected via an incoming arrow contain a token. While replay progresses, the number of tokens are counted [31]:

Let k be the number of different traces from the aggregated log. For each log trace $i(1 \leq i \leq k)$, let n_i be the number of process instances combined in the current trace, m_i the number of missing tokens, r_i the number of remaining tokens, c_i the number of consumed tokens, and p_i the number of tokens produced during the log replay of the current trace. The token-based fitness metric f is defined as follows.

$$f = \frac{1}{2}\left(1 - \frac{\sum_{i=1}^{k} n_i\, m_i}{\sum_{i=1}^{k} n_i\, c_i}\right) + \frac{1}{2}\left(1 - \frac{\sum_{i=1}^{k} n_i\, r_i}{\sum_{i=1}^{k} n_i\, p_i}\right)$$

Trace alignment is established according to the relation between “moves” in the log and “moves” in the model. First, “move on log” represents an event occurring in the log that could not be related to an action in the model. Second, a “move on model” represents the cases where an activity is executed in the model but the log does not contain an event to map to that activity. Third, a “synchronous move” points to an event contained in the log corresponding to an activity executed in the model and vice versa.

Alignment measures the fitness of a trace as a value between 0 and 1. The alignment is maximized with the number of synchronous moves and is minimized with the number of moves on log and moves on model. The value 0 means the poorest fitness between the log and the model while the value 1 corresponds to a perfect fitness, meaning that the alignment only contains synchronous moves. The obtained fitness reflects how the log traces fit to the model and vice versa. In our case, it makes it possible to quantify how much each student’s behaviour is similar to the behaviour followed by the MSS.

In this study, a ProM implementation of the replay technique based on trace alignments is used to check the conformance of each log with an assessment model [32]. The described process is structured as follows.

1. Input: Event log, Assessment model
2. Replay ⟶ Report with fitness for all students
3. Export report + Event log ⟶ Enriched event log
4. Classify students by fitness (Enriched event log)

V. Case Study

In the previous section we described the method to support the assessment in a learning experience based on serious games using process mining techniques. The next step is to implement our proposal in an actual educational process. For this reason, a case study in a higher education context has been carried out. This case study was conducted using an action research method: “action research takes its cues – its questions, puzzles and problems – from the perceptions of practitioners within particular, local practice contexts. It builds descriptions and theories within the practice context itself, and tests them through intervention experiments, that is, action research through experiments that bear the double burden of testing hypotheses and effecting some (putatively) desirable change in the situation” [33].

Specifically, action research in education aims to improve an aspect of the research focus, providing practitioners with new knowledge and understanding of how to improve educational practices or resolve significant problems in learning contexts [34]. This goal can be accomplished by examining actions carried out against the original hypotheses. The theory must solve a practical problem and generate knowledge within our context, the assessment process. To this end, a method based on process mining techniques was proposed. The conducted case study provided an event log to be used as input to test our method. Then, their different stages were applied following the considered research questions. Finally, an analysis of the results for each stage was performed, presenting all detected findings.

A. Study Setup

The case study was conducted through an experiment in the “Databases” course, compulsory for the students in the Computer Science degree program in the University of Cadiz (Spain) during the second semester of their second year. Our experience is focused on the skills related to the design of conceptual data specification through an E/R diagram analysing textual requirements. Therefore, the students need to apply the specific skill for relational databases analysis and design focusing on the learning outcome related to the E/R diagram modelling: knowledge of how to produce a logical and conceptual design of a database. The skill and learning outcomes are included in the course syllabus aligned with the ACM/IEEE Computing Curricula recommendations [35].

The study was carried out in a compulsory workshop where students had to play, individually, a serious game specifically designed for the experience [36]. In all, 110 students participated in the workshop. The video game was developed using the Unity engine and can be run on multiple platforms (Windows, GNU/Linux and MacOs).

The player is challenged to design an E/R diagram according to the provided textual requirements. The proposed exercise inside the game is based on the practical example included in the appendix of [37], a widely used reference for database concepts. The E/R diagram to be designed includes 6 entities, 19 attributes, 7 relations and 14 text boxes to insert relation cardinalities.

B. Overview of the Serious Game

At the beginning of the game, a unique identifier for the player is requested. This identifier is included in the event log and used to relate the performed actions with the specific student. Once the identifier is inserted, the player will be allowed to navigate between different screens through a game menu. In addition, an option to confirm the E/R diagram and exit the game is included.

The map screen is the default screen presented to the player. In this screen, the player must collect the textual requirements for the proposed exercise by clicking on the buildings in the campus. When the player selects a building, a Non Player Character (a class delegate, a lecturer or a department head) provides new textual requirements. The notepad screen includes all the textual requirements collected by the player. Each requirement is composed of one or two sentences where the most important words to be considered for the design of the E/R diagram are highlighted in a different colour.

Last but not least, the player must design the E/R diagram using the provided tools in the editor screen, split into an inventory bar and a work panel if needed. An example of an E/R diagram design is shown in Fig. 2. First, the player can choose an element from the

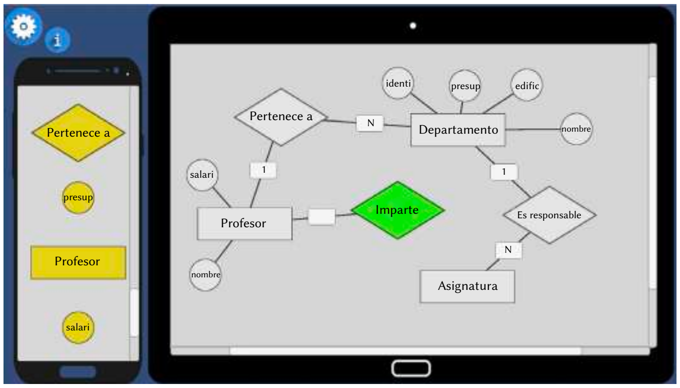


Fig. 2. Example of an E/R diagram being designed in the editor screen (in the Spanish language). The inventory bar is shown on the left and the work zone corresponds to the rest of the screen.

inventory bar and add it to the work panel. Then, the work panel allows the player to organize the elements of the model, relate them, and set their cardinalities (0, 1 or *N*). Finally, the player can easily remove any element of the work panel and add it later as the same or as a different type.

C. Dataset Processing

An event log is created when a player confirms their designed E/R diagram and exits the game. Each event contains: a unique identifier for the player, the performed interaction, a timestamp and a set of additional data according to the type of the interaction. After collecting the resulting event logs for each student, we conducted a processing through the executions of some scripts to join all the event logs into a single dataset. This dataset consisted of 35,931 events. Finally, only those interactions directly related to the design of the E/R diagram were kept, filtering out the other interactions, such as map actions, navigation between screens, etc. An event log with 9,402 events (one per interaction) and 110 traces (one per student) was obtained.

There are 18 types of interactions related to the design of the E/R diagram. These interactions cover the start/end game and the different operations that the player can apply to the different elements of the E/R diagram (entity, attribute and relation): add, delete, link and unlink. In addition, there is an interaction for each value of the cardinality assigned to the union of an entity with a relation (0, 1, *N*) and another to leave it empty.

In-game errors were classified into two categories and seven types, storing the count of each type of error. This classification is based on the usual criteria followed by the course instructor to grade the E/R diagrams designed by the students. Each category represents the level of the error: “major” errors are critical errors and “minor” errors are imprecisions or less important mistakes. On the one hand, 4 types of major errors were considered: fake entity, missing entity, fake relation and missing relation. On the other hand, 3 types of minor errors were considered: fake attribute, missing attribute and fake cardinality. “Fake entities/relations/attributes” refers to E/R diagram elements that should have been designed as another type of element, while “Missing entities/relations/attributes” include those cases where the elements were not designed in the E/R diagram.

In the case of several errors coming from the same origin, only the source error was considered. For instance, if a player missed an entity, only that error was taken into account, ignoring the consequent missed attributes.

The final in-game outcome is automatically calculated in the dataset processing considering the elements (entities, attributes, cardinalities, etc.) of the player’s solution that are similar to those of the diagram solution provided in [37]:

Let s be the number of elements of the E/R diagram designed by the player during the game that are similar to those of the E/R diagram solution, and let t be the total number of elements included in the E/R diagram solution. The final in-game outcome is the ratio r defined as follows.

$$r = \frac{s}{t} * 100$$

This outcome does not represent an usual grade, since we aim for a proper interpretation of the textual requirements and a high level of similarity with the solution. For instance, an E/R diagram with only a 50% similarity was not considered as a valid design.

VI. Discussion of Process Mining Techniques for Skill Assessment

The method discussed in the previous sections was implemented using the processed event log described above as the main input object of ProM. We present and discuss the results of implementing our proposal concerning the research questions previously introduced.

A. Research Question 1 – Student Outcomes According to In-Game Errors

In our first stage, we aim to identify the most decisive in-game errors for the outcomes obtained by two specific student profiles, viz., the MSS and the LSS. A classification of the outcomes based on in-game errors was carried out generating decision trees to obtain clusters of students. An initial use case was conducted to filter the log according to the length of the trace and the in-game outcome. The E/R diagram corresponding to the supplied solution has 6 entities, 19 attributes, 7 relations and 14 cardinality inputs: 46 elements in all. In addition, link events are mandatory to properly build the E/R diagram. Therefore, some traces with few events and a poor final outcome regarding quick tests, where the students only explored the game options, were considered as noise in the event log.

An initial analysis of the use case was conducted to identify and filter the cases considered as noise. First, the in-game outcome of the student was selected as the dependent characteristic and the length of the trace as the single independent characteristic. Then, the event selection filter was not used, maintaining all the events of the log. This analysis provided a decision tree with which to classify the in-game outcome for each trace according to the length of the trace.

The decision tree includes nodes labelled with the corresponding independent characteristic, which was “Trace length” for all nodes in this analysis. The nodes are linked with other nodes or with the generated clusters. All links show the corresponding condition for the value of the independent characteristic. The generated clusters are represented as the leaves of the tree, which are named CL_i (with values of i ranging from 1 to *N*). These clusters are generated by the decision tree algorithm.

The clusters are also labelled with the interval of the values included for the dependent characteristic, with the in-game outcome in this and subsequent analyses. In addition, there were some students who were not classified, because their in-game errors and in-game outcomes did not fulfill the requirements of any cluster. Therefore, not all the outcome intervals or values are covered, because the decision tree only displays the students correctly classified in clusters.

Fig. 3 illustrates a section of the obtained decision tree. In this analysis, we focused on Cluster CL1 because it has the least number of events (≤ 78) and also covers the lowest outcomes: [24.05–51.90). This evidences that this cluster includes the traces of students corresponding to limited game experiences. Cluster CL1 contains 12 traces of students, which were considered as noise and removed to refine the event log. This analysis makes it possible to maintain in the log other students with similar low outcomes but longer traces.

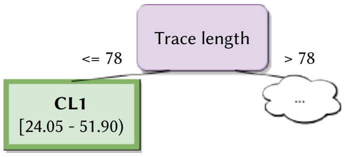


Fig. 3. Decision tree. In-game outcome according to the trace length (CL1).

Once the event log was filtered, how the in-game errors affected the students' outcomes was identified in the next (second) analysis. Again, the in-game outcome of the student was selected as the dependent characteristic. Unlike in the first analysis, several independent characteristics were included, selecting all the possible types of in-game error registered in the event log and listed in the previous section. No event selection filter was used because the log was already filtered according to the results of the first analysis. The obtained decision tree is shown in Fig. 4 and Fig. 5: it has the same structure as that in Fig. 3. In this case, the nodes of the decision tree can have different labels due to the types of error that were selected as the independent characteristics.

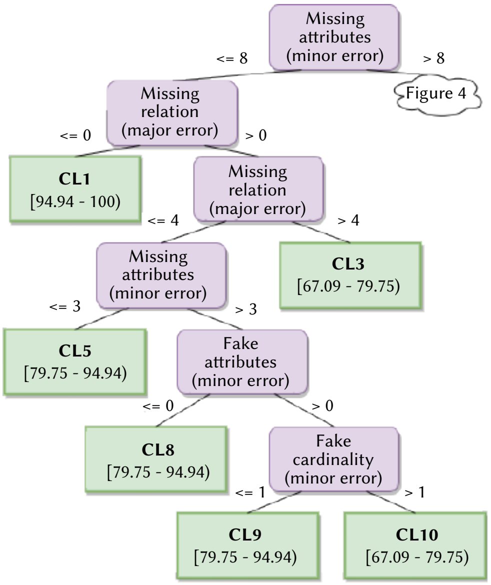

Missing attributes (minor error)
<= 8
> 8
Figure 4
Missing relation (major error)
<= 0
> 0
CL1
[94.94 - 100)
Missing relation (major error)
<= 4
> 4
Missing attributes (minor error)
CL3
[67.09 - 79.75)
<= 3
> 3
CL5
[79.75 - 94.94)
Fake attributes (minor error)
<= 0
> 0
CL8
[79.75 - 94.94)
Fake cardinality (minor error)
<= 1
> 1
CL9
[79.75 - 94.94)
CL10
[67.09 - 79.75)

Fig. 4. Decision tree. In-game outcomes based on in-game errors (Left).

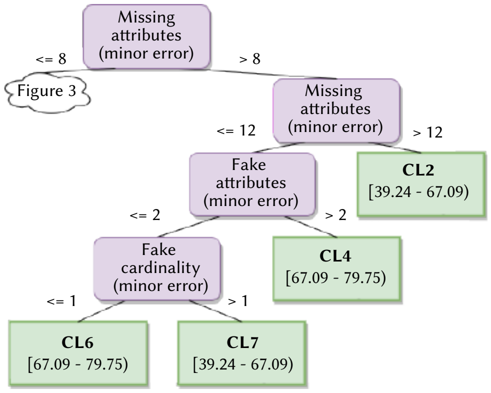

Missing attributes (minor error)
<= 8
> 8
Figure 3
Missing attributes (minor error)
<= 12
> 12
Fake attributes (minor error)
CL2
[39.24 - 67.09)
<= 2
> 2
Fake cardinality (minor error)
CL4
[67.09 - 79.75)
<= 1
> 1
CL6
[67.09 - 79.75)
CL7
[39.24 - 67.09)

Fig. 5. Decision tree. In-game outcomes based on in-game errors (Right).

Cluster CL1 includes the interval with the highest outcomes [94.04–100), with "missing attribute" and "missing relation" their dependent errors. This cluster includes 15 students who missed less than 9 attributes and included all the relations, achieving a final outcome of at least 94.94 out of 100. In addition, students who achieved an outcome equal to 100 were not classified into this cluster because they did not make any in-game errors.

Then, the three clusters with the next high outcomes (CL5, CL8 and CL9) cover the interval [79.75–94.04) and include 24 students in all. First, Cluster CL5 includes up to four missing relations and three missing attributes at most. Then, Cluster CL8 also includes the cases without fake attributes and increases the number of allowed missing attributes to four. Lastly, Cluster CL9 includes the traces with any number of fake attributes, but at most one fake cardinality.

The next interval, corresponding to results lying within the interval [67.09–79.75), is divided into four clusters (CL3, CL4, CL6 and CL10), which include 20 students in all. These clusters depend on the same type of errors that the previous clusters did, but with higher values. Lastly, the interval with the lowest outcomes is [39.24–67.09), which is covered by two clusters (CL2 and CL7) and includes 24 students in all. Cluster CL2 has cases with more than 12 missing attributes while Cluster CL7 depends on more types of error. In addition, two students achieved a similar outcome but they were not classified into these clusters because they made different in-game errors.

Diverse behaviours could have been performed in the clusters obtained due to their wide intervals of outcomes. Therefore, an iteration of the process to focus on specific profiles and obtain more detailed results was conducted. Based on the previous clusters, two sub-logs were created to analyse the successful (CL1) and the less successful students (CL2 and CL7). First, the sub-log corresponding to the successful students includes those students with a final outcome greater than or equal to 94.94, composed of 17 traces and 1,521 events. Then, the sub-log corresponding to less successful students includes students with outcomes less than or equal to 67.09, composed of 26 traces and 2,261 events. The decision trees obtained by these views are shown in Fig. 6 and Fig. 7 and have the same structure as the previous ones.

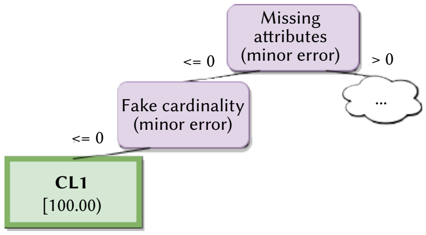

Missing attributes (minor error)
<= 0
> 0
Fake cardinality (minor error)
...
<= 0
CL1
[100.00)

Fig. 6. Decision tree. In-game outcomes for successful students based on in-game errors (CL1).

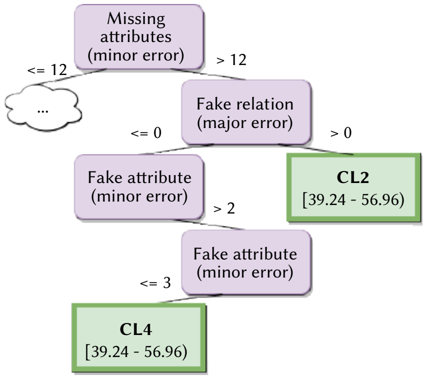

Missing attributes (minor error)
<= 12
> 12
...
Fake relation (major error)
<= 0
> 0
Fake attribute (minor error)
CL2
[39.24 - 56.96)
> 2
Fake attribute (minor error)
<= 3
CL4
[39.24 - 56.96)

Fig. 7. Decision tree. In-game outcomes for less successful students based on in-game errors (CL2 and CL4).

On the one hand, Fig. 6 shows that Cluster CL1 is labelled with [100.00] as it only includes a single value instead of an interval. In this case, this cluster contains the two students who achieved the maximum

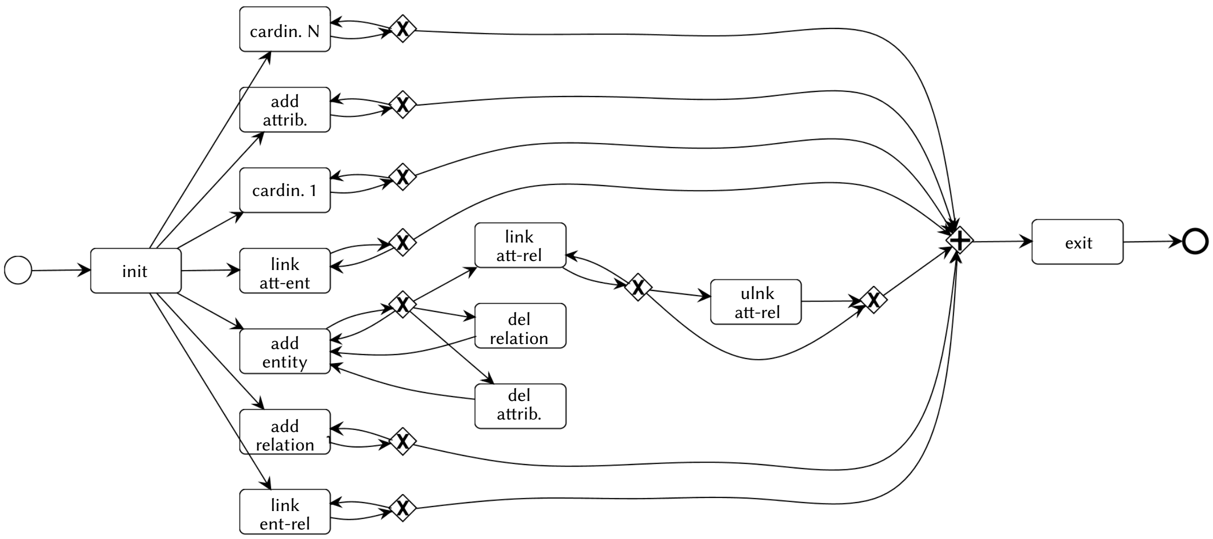


Fig. 8. Most successful students (MSS) including all paths.

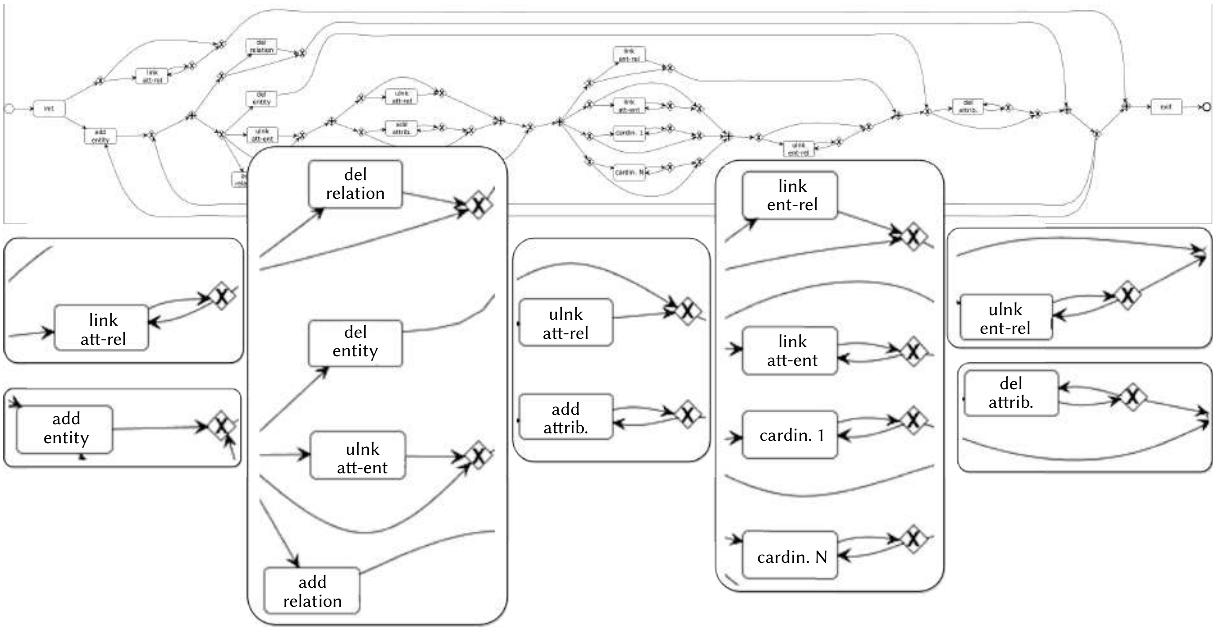


Fig. 9. Least successful students (LSS) including all paths.

outcome in the game. The only errors related to this outcome are missing attributes and fake set cardinalities. Therefore, only those students who properly set the attributes and the cardinalities achieved the maximum outcome. These two students are considered as the MSS. On the other hand, Fig. 7 shows two clusters (CL2 and CL4), which cover the interval [39.24–56.96). In these cases, more than 12 attributes were wrongly set. In addition, these students made fake relations or included up to three fake attributes. The log also includes another student with a similar outcome (51.90) but not classified in the cluster due to different in-game errors. All these students (8) were considered as the LSS.

Based on their in-game errors, the conducted process showed the justification for why a student was included in the respective cluster. According to the results of the previous decision trees, missing attributes was the most decisive type of error to achieve successful outcomes because it is in the top of the three trees. The fake relation type of error also discriminates between the successful and less successful students. Lastly, fake cardinalities and fake attributes were also critical to discriminating between the MSS and the LSS, respectively. These types of error were considered as the most decisive in-game errors for the outcomes achieved by the MSS and the LSS, answering RQ1.

B. Research Question 2 – Behaviour Analysis Between Student Profiles

After identifying the most decisive in-game errors for the MSS and LSS, we aim to apply model discovery techniques to detect similarities and differences in the performance of these student profiles. According to previous results, the corresponding sub-logs from the clusters including MSS and LSS are obtained: [100] and [39.24–56.96), respectively. First, the MSS sub-log includes two traces and 171 events, corresponding to those students who achieved the maximum outcome in the game (100). Second, the LSS sub-log includes 8 traces and 691 events, representing the lowest results after excluding traces considered as noise due to a scarce trace length.

Firstly, the event sub-logs were loaded in ProM to carry out the model discovery. In Business Process Model Notation (BPMN), empty circles represent the starting and the ending states. Then, all the nodes are

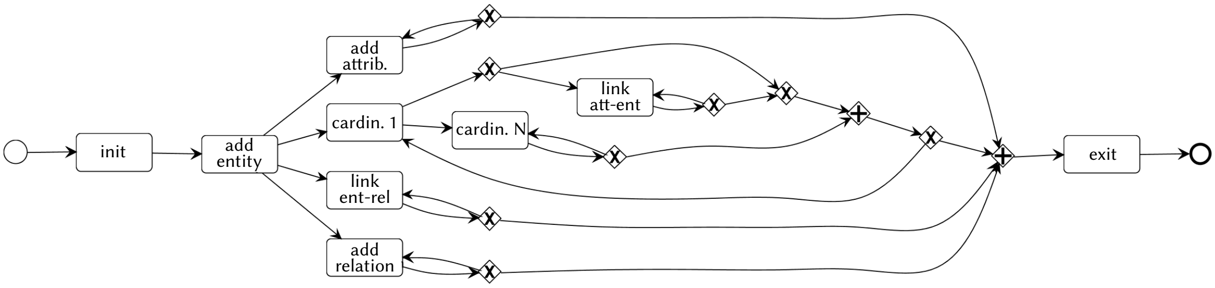


Fig. 10. Most successful students (MSS) filtering out less frequent paths

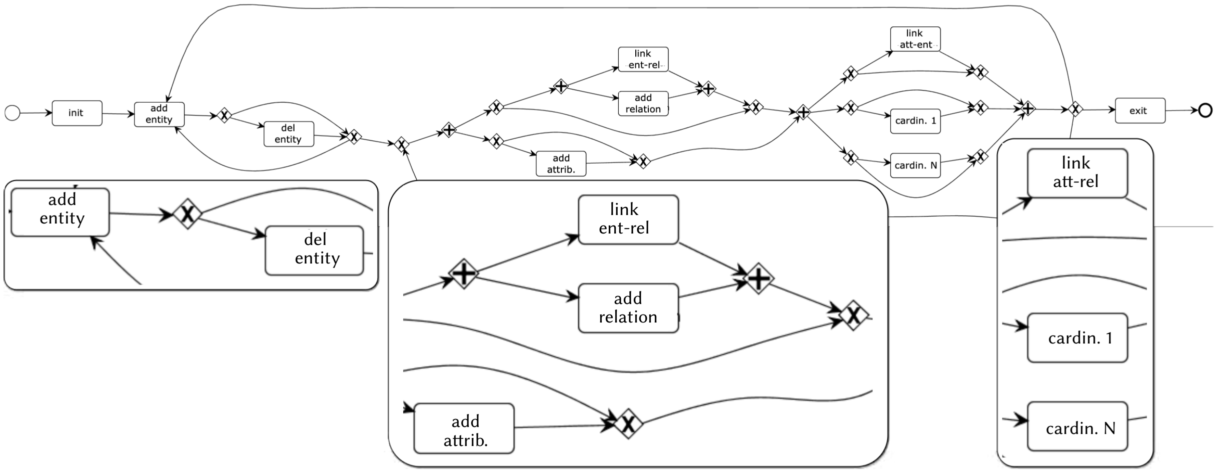


Fig. 11. Least successful students (LSS) filtering out less frequent paths.

labelled with an event action, which in this context corresponds to one type of player interaction. These nodes are linked by directed arrows to represent the sequentiality. In addition, two types of diamonds are present to control the flow. Diamonds with a "+" represents a join for input paths and a parallel split of output paths to follow all the subsequent gateways. Diamonds with a "X" corresponds to decision points where the path bifurcates in case of having multiple outputs. The names of the events have been shortened to improve the visualization.

In the first iteration, models for the MSS and the LSS profiles were generated by using IM, as shown in Fig. 8 and Fig. 9, respectively. The wide variety of sequences are reflected in the model for MSS (Fig. 8) through numerous loops, enabling practically arbitrary executions of activities. In addition, all the types of event included in the sub-log (13), even the least frequent ones, are present in the model. The visual inspection was supported by checking the events included in the sub-log, which confirmed that any traces started with interactions such as link-ent, link-rel, cardin. 1 or cardin. *N*; but the model made this behaviour possible. Therefore, the resulting model is too imprecise, being underfit and enabling too many behaviours, thus it does not provide a good reference for the students' behaviour. Similar problems are evidenced in the model for LSS (Fig. 9). Although this model does not present as many arbitrary loops as the model for MSS, it presents a too complex structure and too many paths to skip the majority of activities of the process. Finally, it also presents all the types of events included in the sub-log (16), which can result in the inclusion of infrequent event activities.

Considering these issues, IMi was implemented in the next iterations for the MSS and the LSS profiles with different incremental values for the noise threshold. The use of IMi makes it possible to select a noise threshold from 0.00 to 1.00, where setting 0.00 guarantees a perfect log fitness.

Searching for a balance between precision and generalization ability, less frequent paths were filtered by setting a 20% noise threshold. The resulting models for MSS and LSS are shown in Fig. 10 and Fig. 11, respectively. Compared with the previous ones, both models are more suitable. These models are more precise and do not include the less frequent types of events: the MSS model includes 9 instead of 13 and the LSS model includes 10 instead of 16. Therefore, infrequent paths were also filtered out. In addition, the models have a simpler structure, easier to understand and more significant.

The discovered models for the MSS and the LSS profiles still have several loops even though the infrequent paths were filtered out. In essence, the design of an E/R diagram is carried out through multiple loops for the different types of actions, such as adding elements and creating links. Therefore, we assume that the occurrence of these loops is a consequence of the iterative nature of the analysed process.

Significant differences in the structure between the MSS and LSS models can be seen (Fig. 10 and Fig. 11, respectively). One the one hand, the MSS model includes the add entity interaction and several small loops for the rest of the interactions, so many paths are possible. However, the setting of the cardinalities and linking attributes are in the same path. This sequentiality could suggest that these types of interactions were usually done in sequence. This is a recommended pattern in the design of E/R diagrams, as both aspects are deeply interrelated. On the other hand, the LSS model has two big loops and a more linear structure, with three subprocesses. First, there are the add and delete entity interactions. This similarity with the MSS model makes sense due to fact that the entities can be considered as the starting point of the design of an E/R diagram. Second, attributes and relations are added. In addition, this subprocess includes the link between entities and relations. Third, cardinalities are set and attributes with entities are linked. This behaviour is similar to that performed by the MSS because these interactions were grouped as well. In general, this structure suggests a common heterogeneity in the behaviour performed by the LSS but with differences because of the multiple bifurcations the model presents.

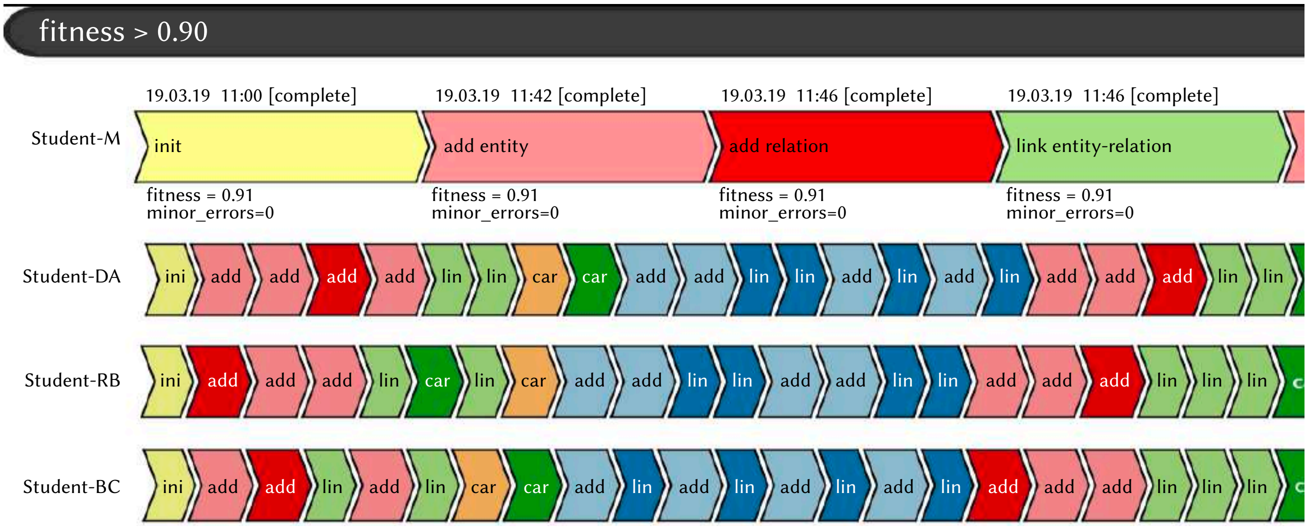

fitness > 0.90
Student-M
19.03.19 11:00 [complete]
init
fitness = 0.91
minor_errors=0
19.03.19 11:42 [complete]
add entity
fitness = 0.91
minor_errors=0
19.03.19 11:46 [complete]
add relation
fitness = 0.91
minor_errors=0
19.03.19 11:46 [complete]
link entity-relation
fitness = 0.91
minor_errors=0
Student-DA
ini add add add add lin lin car car add add lin lin add lin add lin add add add lin lin
Student-RB
ini add add add lin car lin car add add lin lin add add lin lin add add add lin lin lin c
Student-BC
ini add add lin add lin car car add lin add lin add lin add lin add add add lin lin lin c

Fig. 12. Traces of students who obtained a fitness > 0:90.

The MSS model only includes interactions that add value to the design of the E/R: there are no removing interactions such as deleting elements or unlinks. Although its corresponding general model (Fig. 8) included some of these types of interactions, they were filtered out once the infrequent paths were removed. However, the model for LSS (Fig. 11) still includes one interaction of this type after filtering out the infrequent paths: "delete entity". This behaviour evidenced an indecision at the beginning of the process. The creation of entities is the most basic step in the design of an E/R diagram, so the performance also evidences some limitations to the use of design skills. In addition, the MSS model shows the students added all the entities at the beginning while the LSS model includes loops to return to this point and open the path to new deletions of entities. This difference in the behaviour could be evidence that the indecision continued in later stages of the design process.

The lack of other removal interactions, such as deletes or unlinks, in the LSS model could suggest that the hesitations were only focused on interpreting the requirements to detect what elements should be designed as entities. This is supported by checking the event log, observing that less than 5% of the events performed by the LSS were removal interactions other than delete entities. However, as discussed in Section V.B, the serious game only considered the source error when several errors come from the same origin. Therefore, the LSS could have made consequent errors and had hesitations interpreting the whole of the requirements.

Except for "delete entity", the models for MSS and LSS include the same types of interaction. Although their general models (Fig. 8 and Fig. 9) present more variations between them, the majority of these variations were removed from the models when filtering out infrequent paths. This similarity can be evidence that the LSS usually avoided infrequent or not required interactions, such as "link attribute-relation" and "cardinality 0".

According to the discovered models, the results show that there are some similarities and differences in the performance between the MSS and the LSS. All these similarities and differences answer affirmatively the RQ2, in that both profiles follow similar behaviours but with differences in some key aspects.

C. Research Question 3 – Student Classification According to Performance

The aim of this work is to support student assessment through a behaviour comparison using an assessment model or profile. This comparison was conducted applying a conformance checking to replay the traces of all students to the MSS model discovered in the previous subsection. This process provided the fitness for each trace according to the MSS model, quantifying the similarity between each student and the MSS behaviour. In this process, the MSS model previously discovered after applying IMi to ignore infrequent events and paths was used.

First, the whole event log and the MSS model were loaded in ProM to be used as input in a plugin called "Replay a Log on Petri Net for Conformance Analysis", an implementation that can use several algorithms to replay traces over a model. This plugin was selected because it yields detailed results, providing specific fitness values for each trace and making it possible to export them in an external file. In this study, we used the Dijkstra and ILP algorithms, which provided similar results. This plugin was used to replay the log and calculate the fitness values according to the alignment between the log and the model. As we discussed during the presentation of the conformance checking stage of our method in Section IV.E, this process provides a conformance checking report that can be exported and processed to extract the fitness values. Then, the original event log was enriched by including the corresponding fitness in each student's trace. Students with similar in-game outcomes could have carried out different behaviours, so the fitness could be different as well. Finally, the enriched event log was loaded in ProM to be explored and classify the traces.

Aiming to compare the in-game outcomes and behaviours (fitness), the average value for the fitness of those who share the same in-game outcome was calculated. Finally, the in-game outcome and fitness were associated, employing the Pearson correlation coefficient, obtaining 0.75. We can affirm that the students' behaviour and their achieved in-game outcomes have a reliable correlation in the context of this case study.

ProM enables filtering traces according to the fields included in the event log. Fig. 12 shows all the students' traces with a fitness higher than 0.90. The different traces are shown in independent lines and include the sequence of events. The traces also contain the unique identificator for each student, anonymized as "Student-" plus one or more capital letters. This information was provided to the students who fulfilled the filter requirements. Each trace is composed of small boxes that correspond to the events of the corresponding student. The events are labelled with the shortened name and are displayed using different colours for each type of event. For instance, "init" is shortened to "ini" and displayed in yellow.

The conformance checking stage of our method provides an interactive chart to explore the traces and events. In Fig. 12, we selected the first student (Student-M) to expand their events and show the specific dates and times of each event. A sample of the included fields is displayed under the trace but all the fields with their types and values can be explored after the traces. For instance, "Student-M" is part of the MSS students because that student achieved an in-game outcome of 100, that is, no errors were made.

This process can be iterated using different fitness values and obtaining the corresponding students. The traces of students with a fitness between 0.90 and 0.80 were obtained. This provided 82 traces, meaning a high ratio of the traces included in the whole event log (74.54%). This evidences a high similarity in the behaviour of the majority of the students. To obtain more detailed results, a shorter range for the fitness value was used for the filter (0.90 and 0.89). This filter provided 10 traces. After exploring these data, we found that 8 of these students achieved in-game outcomes higher than 89.00, showing a positive correlation with their fitness. In addition, the remaining two students (Student-FD and Student-QC) achieved unusual in-game outcomes (51.90 and 67.09, respectively). Although those two students achieved lower results, they showed a similar behaviour to the others with high in-game outcomes. Therefore, unlike assessment methods based only on the obtained grade, this assessment method considered behavioural evidence to detect how the students applied their skills.

Beyond the Pearson correlation coeÿcient previously calculated (0.75), the course's supervisor provided positive feedback about being supported by the proposed assessment method. This provides objective evidence about the behaviour followed by the students during the experience. The obtained fitness value allowed the supervisor to detect students who, despite carrying out proper behaviour during the process, did not have this reflected in the grade obtained. Multiple factors could be involved in this result, i.e. good skills about E/R designing but a lack of requirements analysis. Therefore, the supervisor used this assessment method to optimize the revision process, revising only specific cases where discrepancies between grades and behaviour were found.

In accordance with the conducted conformance checking, the replay report provided a fitness value for each student trace. This fitness quantifies the similarity between one student's behaviour and the assessment model. It was incorporated in the event log as an additional field, so all the students could be classified according to their performance during the game. Therefore, the employed filters provided a scalable support for the students' assessment through a behaviour comparison using assessment models, thus answering RQ3.

D. Threats to Validity

After using the proposed method in a case study, it is essential to identify potential threats to its validity that may occur in the development of the study. This subsection of the paper discusses possible construct threats to the validity of the experiment and our proposal to measure/mitigate them.

Some of the configurations used during the process could be coupled to the analysed dataset, such as the selected noise threshold to filter out infrequent paths. This issue was considered as a challenge to internal validity. All the analysed data to replicate the experiment and additional figures to detail the results are available in an open multimedia repository [36]. This information ensures the reliability of this study and our results.

The limitations of assessment processes discussed in the paper were detected in studies where the assessment of the learning process was focused on the acquisition of skills. As we have not reviewed any learning processes with assessment focused on getting knowledge, we can not evaluate if our method can be adapted or not for those learning processes. This limitation was considered as a challenge to the external validity. A literature review, additional modeling and empirical research are necessary to confirm if learning processes with assessment focused on getting knowledge present the same limitations.

Our proposal has two requirements. First, the student has to resolve the game by applying one or more specific skills and, second, the game has to provide logs with relevant information for the assessment. The adaptation of the method in other skill assessment problem is dependent on how well the game helps/requests the development of the skill from the students, and the quantity and quality of the information stored in the logs. Once a serious game provides logs with relevant information to be considered in the skill assessment, the proposed method allows to easily analyze in detail thousands of events produced during the game experience. These requirements were considered as a challenge to the external validity.

Although the proposed method is generic and can be used in diverse contexts, the results of this paper are limited to the scope of the conducted case study. All the configurations employed during the process depend on the skill to be assessed and the information stored in the dataset. For instance, if we want to assess a time-focused skill in resolving a task assignment, we could obtain the models considering the time invested in the game experience instead of the in-game errors. Additionally, the dataset could be filtered according to a minimum in-game outcome to only consider those students who successfully passed the assignment.

In order to generalise the findings to other processes, a replication of the experiment in other learning experiences is needed.

Regarding the use of our method in other database courses, the assessed skill in this paper is included in the course syllabus aligned with the ACM/IEEE Computing Curricula recommendations [35]. It is a widely used reference for higher education computer science programs, so the proposed method should adjust as well to a database course other than the one taken as source for this particular study. Since the presented case study was conducted in a single database course, the limitations previously discussed should be considered as well.

The assessment was supported in an experience based on a serious games, through process mining techniques. The learning experience aims to assess a skill for the analysis and design of relational databases, focussing on the design of E/R diagrams. However, this kind of process has an intrinsically iterative essence because the same events are performed multiple times during the game. This feature resulted in process models with multiple loops and bifurcations even after filtering out infrequent paths. Addressing this iterative essence, applying process mining techniques, was a challenge to the construction validity.

VII. Conclusions

As learning processes are focused on the acquisition of skills, students must be assessed according to their level of proficiency in these skills. In this paper, we aim at supporting a solution for skill assessment through behaviour comparison, using assessment models. In order to validate the proposal, a case study was conducted in a course on databases, part of a Computer Science degree program. More than 100 students had to apply database analysis skills by designing an E/R diagram during a serious game. In all, the interactions of the students provided 35,931 events, which were processed and refined to 9,402 events. The main research question is whether process mining techniques can support scalable assessment in a learning experience based on serious games. In order to answer this question, it was divided into three research questions.

In the first question (RQ1), we aimed at identifying the most decisive in-game errors for specific student profiles: the most successfull students (MSS) and the least successfull students (LSS). A clustering of dynamic behaviour through decision trees was employed to generate clusters and classify the students. How in-game errors affect the students' outcomes was identified in an analysis of a use case. First, having missing attributes was the type of error that was most decisive in failing to achieve successful outcomes, because it is in the top of the majority of the trees obtained. Then, another type of error was also critical for differentiating between successful and less successful students (fake relations). Additional errors differentiated the MSS from the rest of the successful students (fake cardinalities) and differentiated the LSS from the other students with poor results (fake attributes).

In the second question (RQ2), we aimed at detecting similarities and differences between the performances of specific student profiles. Based on previous analysis, models for the MSS and the LSS were obtained applying model discovery techniques through inductive mining. The results showed some similarities and differences in the performances of both student profiles. First, we detected similarities in the type of interactions, iterative performance, starting point for the E/R design, and subprocess for specific interactions (cardinalities and link attribute-entity). Second, the models had differences in their main structure, different types of loops, differences in the occurrence of bifurcations and of interactions to delete entities in the LSS model.

In the third question (RQ3), we aimed at applying process mining techniques to assess the students according to their performance during the game. We proposed integrating the assessment process based on the comparison of the performance using assessment models or profiles. Therefore, the previously discovered model for the MSS was used as the assessment model to replay the whole event log for a conformance checking. Conformance checking supported the inferences from a visual inspection of the process models. The replay provided a fitness value for each student trace, measuring the similarity of the behaviour of each student to the MSS. The obtained fitness and the in-game outcomes presented a reliable and positive correlation in this context. Finally, the fitness was used to enrich the event log and classify students according to their performance.

The proposed method in this paper has several similarities with other automated assessment methods reviewed in this paper. Although there is no formal relation between indicators and evidence, a visual inspection of the discovered models was made that focused on detecting patterns and evidence of the applied skills during the game. As a stealth assessment method, we also collected data during the game to gather evidence through a quiet process, removing test anxiety as much as possible. More specifically, in this case study, the design of an E/R was an iterative process that can produce a large quantity of interactions for which a manual analysis is not feasible. Process mining techniques proved to be a suitable solution to handle them.

Unlike other data-centred automated assessment methods, process mining is a mixed approach, lying between data-centred and process-centric. The proposed method can model students' behaviour and discover knowledge from event data. First, this mixed approach makes it possible to use event data for detecting the most decisive in-game errors and how they determine the students' outcomes. Second, the process-centred techniques are used to model students' behaviour and compare their performance.

In addition, other automated assessment methods are usually focused on a formative aim, specially in intelligent tutoring systems. We aimed at a summative assessment purpose, but focusing on the behaviour performed during the complete game experience instead of only the final results. This additional feedback allowed the supervisor to detect students who, despite carrying out a proper behaviour during the process, did not have this reflected in the grade obtained.

The scope of the validation for the proposed method is limited to a single skill of "analysis and design of relational databases". However, the method provides techniques to assess other skills because it makes it possible to use several assessment models or profiles, i.e. students with the highest outcomes in specific features. Therefore, independent assessment models can be defined as long as the skills need to be assessed independently.

Although we used an assessment model discovered from the analysed data set (i.e. students with the highest in-game outcomes), this model could have been previously defined by an expert in the area of application (i.e. databases), for instance, by using a model that represents the expert's behaviour, or by using external tools.

The fitness values obtained in the replay represent the indicators to be considered to assess the behaviour of students during the game. Fitness measures the gap between the behaviour represented by the assessment model and the behaviour of each student. An implication to take into account so as to put the method into practice is that fitness values have to be exported from the conformance checking report. Using the fitness, filters were applied to detect the level of similarity that students achieved with the "expert" behaviour. This fitness value could also be used as an input in additional methods.

This paper provides a methodological contribution to the use of process mining techniques to support skill assessment in serious games. Other skill assessments based on process mining techniques for learning experiences based on serious games have not been identified in the literature. In addition, the serious game used in the case study is a software contribution specifically developed and aligned with the assessed skill: the design of a conceptual data specification through an E/R diagram analysing textual requirements. Other serious games with the same purpose have not been identified in the literature.

As the main conclusion, process mining techniques can support a scalable assessment method in learning experiences based on serious games. Applying the process mining model discovery was suitable for analysing the behaviour of students in sequential E/R diagram modelling processes. The tools used provided an automated support to assessing the developed skills during the gameplay. Therefore, we consider the previous evidence as positive to answer the research question of this study.

Regarding future work, we have in view extending the developed serious game with new features and improvements to enable the assessment of other skills from the "Databases" course, beyond the design of a conceptual data specification through an E/R diagram analysing textual requirements. Another line of research would involve applying the method to different types of learning experiences to study learning processes in different contexts. Finally, an additional line of research would be focused on the teaching-learning process, comparing the results obtained by the students with the teacher's assessment.

Acknowledgments

This paper is part of the R&D project CRÊPES (ref. PID2020-115844RB-I00), funded by MCIN / AEI doi:10.13039/501100011033/. We thank Gregorio Rodríguez Gómez for his advice on assessment processes background.

References

[1] D. Djaouti, J. Alvarez, J.-P. Jessel, O. Rampnoux, "Origins of Serious Games," in *Serious Games and Edutainment Applications*, M. Ma, A. Oikonomou, L. C. Jain Eds., London: Springer London, 2011, pp. 25–43, doi: 10.1007/978-1-4471-2161-9_3.

[2] T. M. Connolly, E. A. Boyle, E. Macarthur, T. Hainey, J. M. Boyle, "A systematic literature review of empirical evidence on computer games

and serious games," *Computers & Education*, vol. 59, pp. 661–686, 2012, doi: 10.1016/j.compedu.2012.03.004.

[3] J. A. Caballero-Hernández, M. Palomo-Duarte, J. M. Dodero, "Skill assessment in learning experiences based on serious games: A Systematic Mapping Study," *Computers and Education*, vol. 113, pp. 42–60, oct 2017, doi: 10.1016/j.compedu.2017.05.008.

[4] G. Siemens, S. Dawson, G. Lynch, "Improving the Quality and Productivity of the Higher Education Sector Policy and Strategy for Systems-Level Deployment of Learning Analytics," Society for Learning Analytics Research, 2013.

[5] W. van der Aalst, A. Adriansyah, A. K. A. De Medeiros, F. Arcieri, T. Baier, T. Blickle, J. C. Bose, P. Van Den Brand, R. Brandtjen, J. Buijs, A. Burattin, J. Carmona, M. Castellanos, J. Claes, J. Cook, N. Costantini, F. Curbera, E. Damiani, M. De Leoni, ... M. Wynn, "Process mining manifesto," in *Lecture Notes in Business Information Processing*, vol. 99 LNBIP, 2012, pp. 169–194, Springer Verlag.

[6] T. Hainey, T. M. Connolly, Y. Chaudy, E. Boyle, R. Beeby, M. Soflano, "Assessment integration in serious games," in *Psychology, Pedagogy, and Assessment in Serious Games*, IGI Global, nov 2013, pp. 317–341, doi: 10.4018/978-1- 4666-4773-2.ch015.

[7] M. Allen, *Assessing academic programs in higher education.* Bolton, MA: Anker, 2004.

[8] V. J. Shute, "Stealth Assessment in Computer-Based Games To Support Learning," in *Computer Games and Instruction*, S. Tobias, J. D. Fletcher Eds., Cambridge: MIT Press, 2011, ch. 20, pp. 503–524.

[9] R. J. Mislevy, G. D. Haertel, "Implications of evidence- centered design for educational testing," *Educational Measurement: Issues and Practice*, vol. 25, no. 4, pp. 6–20, 2006, doi: https://doi.org/10.1111/j.1745-3992.2006.00075.x.

[10] R. J. Mislevy, R. G. Almond, J. F. Lukas, "A brief introduction to evidence-centered design," *ETS Research Report Series*, vol. 2003, no. 1, pp. i–29, 2003, doi: 10.1002/j.2333-8504.2003.tb01908.x.

[11] V. J. Shute, M. Ventura, M. Bauer, D. Zapata-Rivera, "Melding the power of serious games and embedded assessment to monitor and foster learning," in *Serious games: Mechanisms and effects*, U. Ritterfeld, M. J. Cody, P. Vorderer Eds., Routledge, 2009, pp. 295–321.

[12] V. J. Shute, Y. J. Kim, "Formative and Stealth Assessment," in *Handbook of Research on Educational Communications and Technology*, J. M. Spector, M. D. Merrill, J. Elen, M. J. Bishop Eds., New York, NY: Springer New York, 2014, pp. 311–321, doi: 10.1007/978-1-4614-3185-5_25.

[13] A. Mitrovic, M. Mayo, P. Suraweera, B. Martin, "Constraint-Based Tutors: A Success Story," in *Engineering of Intelligent Systems*, Berlin, Heidelberg, 2001, pp. 931–940, Springer Berlin Heidelberg.

[14] P. Suraweera, A. Mitrovic, "An Intelligent Tutoring System for Entity Relationship Modelling," *International Journal of Artificial Intelligence in Education*, vol. 14, pp. 375–417, 2004.

[15] L. Zhuhadar, S. Marklin, E. Thrasher, M. D. Lytras, "Is there a gender difference in interacting with intelligent tutoring system? Can Bayesian Knowledge Tracing and Learning Curve Analysis Models answer this question?," *Computers in Human Behavior*, vol. 61, pp. 198–204, 2016, doi: https://doi.org/10.1016/j.chb.2016.02.073.

[16] M. V. Yudelson, K. R. Koedinger, G. J. Gordon, "Individualized Bayesian Knowledge Tracing Models," in *Artificial Intelligence in Education*, Berlin, Heidelberg, 2013, pp. 171–180, Springer Berlin Heidelberg.

[17] W. M. P. van der Aalst, *Process Mining Data Science in Action.* Berlin Heidelberg: Springer, 2nd ed. ed., 2016.

[18] K. Engelmann, M. Bannert, "Analyzing temporal data for understanding the learning process induced by metacognitive prompts," *Learning and Instruction*, p. 101205, 2019, doi: https://doi.org/10.1016/j.learninstruc.2019.05.002.

[19] H. A. V. D. Berg, "Occam's razor: From ockham's via moderna to modern data science," *Science Progress*, vol. 101, no. 3, pp. 261–272, 2018, doi: 10.3184/003685018X15295002645082.

[20] M. de Leoni, W. M. van der Aalst, M. Dees, "A general process mining framework for correlating, predicting and clustering dynamic behavior based on event logs," *Information Systems*, vol. 56, no. July, pp. 235–257, 2016, doi: 10.1016/j.is.2015.07.003.

[21] A. Bogarín, R. Cerezo, C. Romero, "A survey on educational process mining," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, p. e1230, jan 2018, doi: 10.1002/widm.1230.

[22] A. Bolt, M. de Leoni, W. M. P. van der Aalst, P. Gorissen, "Exploiting Process Cubes, Analytic Workflows and Process Mining for Business Process Reporting: A Case Study in Education," in *International Symposium on Data-driven Process Discovery and Analysis (SIMPDA)*, Vienna, Austria, 2015, pp. 33–47.

[23] J. A. Caballero-Hernández, A. Balderas, M. Palomo- Duarte, P. Delatorre, A. J. Reinoso, J. M. Dodero, "Teamwork assessment in collaborative projects through process mining techniques," *International journal of engineering education*, vol. 36, no. 1, pp. 470– 482, 2020.

[24] P. Reimann, "Time is precious: Variable- and event- centred approaches to process analysis in CSCL research," *International Journal of Computer-Supported Collaborative Learning*, vol. 4, no. 3, pp. 239–257, 2009, doi: 10.1007/s11412-009-9070-z.

[25] M. Bannert, P. Reimann, C. Sonnenberg, "Process mining techniques for analysing patterns and strategies in students' self-regulated learning," *Metacognition and Learning*, vol. 9, no. 2, pp. 161–185, 2014, doi: 10.1007/s11409-013-9107-6.

[26] P. Reimann, L. Markauskaite, M. Bannert, "e-Research and learning theory: What do sequence and process mining methods contribute?," *British Journal of Educational Technology*, vol. 45, no. 3, pp. 528–540, 2014, doi: 10.1111/bjet.12146.

[27] H. M. W. Verbeek, J. C. A. M. Buijs, B. F. Van Dongen, W. M. van der Aalst, "ProM: The Process Mining Toolkit," in *International Conference on Business Process Management Demonstration Track*, Hoboken, New Jersey, 2010, pp. 34–39.

[28] V. U. Kumar, A. Krishna, P. Neelakanteswara, C. Z. Basha, "Advanced prediction of performance of a student in an university using machine learning techniques," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 2020, pp. 121–126.

[29] S. A. Alasadi, W. S. Bhaya, "Review of data preprocessing techniques in data mining," *Journal of Engineering and Applied Sciences*, vol. 12, no. 16, pp. 4102–4107, 2017, doi: 10.3923/jeasci.2017.4102.4107.

[30] S. J. J. Leemans, D. Fahland, W. M. P. van der Aalst, "Discovering Block-Structured Process Models from Event Logs Containing Infrequent Behaviour," in *Business Process Management Workshops. BPM 2013. Lecture Notes in Business Information Processing, vol 171.*, N. Lohmann, M. Song, P. Wohed Eds., Springer, Cham, 2014, pp. 66–78, doi: 10.1007/978-3-319-06257-0_6.

[31] A. Rozinat, W. M. P. van der Aalst, "Conformance checking of processes based on monitoring real behavior," *Information Systems*, vol. 33, no. 1, pp. 64–95, 2008, doi: https://doi.org/10.1016/j.is.2007.07.001.

[32] W. van der Aalst, A. Adriansyah, B. van Dongen, "Replaying history on process models for conformance checking and performance analysis," *WIREs Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 182– 192, 2012, doi: 10.1002/widm.1045.

[33] C. Argyris, D. A. Schön, "Participatory action research and action science compared: A commentary," *American Behavioral Scientist*, vol. 32, no. 5, pp. 612–623, 1989.

[34] G. E. Mills, *Action research: A guide for the teacher researcher.* Boston: Pearson, 4th ed., 2011.

[35] ACM, IEEE, "Computer Engineering Curricula 2016," ACM, IEEE, 2016.

[36] J. A. Caballero-Hernández, "Supporting skill assessment in learning experiences based on serious games through process mining techniques," 2020. doi: 10.6084/m9.figshare.c.4916412.

[37] A. Silberschatz, H. F. Korth, S. Sudarshan, *Database system concepts.* New York: McGraw-Hill, 6th ed. ed., 2011.

Juan Antonio Caballero-Hernández

Juan Antonio Caballero-Hernández received his MSc degree in computer science and his PhD degree from the University of Cadiz, Spain. His main research interest is focused on learning experiences based on serious games. Beyond the academic environment, he has worked in different positions in IT, such as web development and managing teams.

Manuel Palomo-Duarte

Manuel Palomo-Duarte received his MSc degree in computer science from the University of Seville and his PhD degree from the University of Cadiz, where he works as an Associate Professor. He is the author of more than 20 papers published in indexed journals and more than 30 contributions to international academic conferences about learning technologies, serious games and the collaborative Web.

Juan Manuel Dodero

Juan Manuel Dodero is Full Professor of Computer Science in the University of Cadiz, Spain. He has a Computer Science degree from the Polytechnic University of Madrid and a PhD degree from the Carlos III University of Madrid. His main research interests include Web science and engineering and technology-enhanced learning, fields in which he has co-authored numerous research papers in international journals and conferences.

Dragan Gašević

Dragan Gašević is Distinguished Professor of Learning Analytics in the Faculty of Information Technology and Director of the Centre for Learning Analytics at Monash University. He served as the president (2015–2017) of the Society for Learning Analytics Research (SoLAR). A computer scientist by training and skills, Dragan considers himself a learning analyst who develops computational methods that can shape next-generation learning technologies and advance our understanding of self-regulated and collaborative learning. Dragan is a (co-) author of numerous research papers and books and a frequent keynote speaker.

Requirements for User Experience Management - A Tertiary Study

Andreas Hinderks[1], Francisco José Domínguez Mayo[1], María José Escalona[1], Jörg Thomaschewski[2] *

[1] University of Seville (Spain)
[2] University of Applied Science Emden/Leer (Germany)

* Corresponding author. andreas@hinderks.org (A. Hinderks), fjdominguez@us.es (F. J. Domínguez Mayo), mjescalona@us.es (M. J. Escalona), joerg.thomaschewski@hs-emden-leer.de (J. Thomaschewski).

Received 25 March 2021 | Accepted 12 April 2023 | Early Access 30 January 2024

unir LA UNIVERSIDAD EN INTERNET

Abstract

Today's users expect to be able to interact with the products they own without much effort and also want to be excited about them. The development of a positive user experience must therefore be managed. We understand management in general as a combination of a goal, a strategy, and resources. When applied to UX, user experience management consists of a UX goal, a UX strategy, and UX resources. We conducted a tertiary study and examined the current state of existing literature regarding possible requirements. We want to figure out, what requirements can be derived from the literature reviews with the focus on UX and agile development. In total, we were able to identify and analyse 16 studies. After analysing the studies in detail, we identified different requirements for UX management. In summary, we identified 13 requirements. The most frequently mentioned requirements were prototypes and UX/usability evaluation. Communication between UX professionals and developers was identified as a major improvement in the software development process. In summary, we were able to identify requirements for UX management of People/Social, Technology/Artifacts, and Process/Practice. However, we could not identify requirements for UX management that enabled the development and achievement of a UX goal.

Keywords

Agile Methods, Tertiary Study, User Experience, UX Management.

DOI: 10.9781/ijimai.2024.01.004

I. Introduction

A successful product is characterized by its ability to generate a high level of satisfaction among users. Today's user expects to be able to interact with a product without much effort. The user also wants to be excited about interacting with the product. These hedonic interaction qualities must be taken into account in product development. They are usually characterized by the fact that they are not directly goal-directed [1]. In summary, the user wants to have a positive user experience while interacting with the product or service.

ISO 9241-210 [2] defines the term user experience among other terms. It is defined as 'a person's perceptions and responses that result from the use or anticipated use of a product, system or service'. The user experience is thus considered as a holistic concept. Any kind of emotional, cognitive or physical response, be it concrete or even suspected, is considered. The definition of user experience covers the period of time before, during and after the interaction with the product.

Agile methods have been established since the publication of the first version of the Scrum Guide [3]. Software development companies use agile methods (e.g. Scrum [4], Kanban [5], or Extreme Programming (XP) [6]) to develop products or services more efficiently [7]. The iterative approach makes it possible to react to new requirements or changes [8]. This distinguishes agile methods significantly from classic process models such as waterfall. By performing retrospectives [4] at the end of an iteration, both product quality and agile process quality can be improved.

To develop the best possible product with great user experience, it is essential to have the right management in place in terms of UX. To the best of our knowledge, there is no approved definition of UX management in literature. There is also no common understanding of what UX management is or how to apply it.

In this paper, we conducted a tertiary study to discover what requirements for UX management could be derived based on the literature. The research question for the tertiary study is:

RQ: What requirements can be derived from literature reviews for User Experience Management with the focus on agile development?

This paper is structured as follows: Section II briefly summarizes the related work and presents gap analysis. Section III presents the review method including search strategy, selection process, and data extraction. Section IV outlines the results and key findings of our study

Please cite this article as:
A. Hinderks, F. J. Domínguez Mayo, M. J. Escalona, J. Thomaschewski, "Requirements for User Experience Management - A Tertiary Study", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 6, pp. 160-167, 2024, http://dx.doi.org/10.9781/ijimai.2024.01.004

as well as the answers to our research question. Section V discusses the meaning of the findings and the limitations of our study. The paper ends with Section VI, with conclusions and ideas for future work.

II. Related Work

In the beginning, we did informal research on UX management or related terms. We conducted the informal research with Science Direct, Springer Link, IEEEXplore, Scopus, and ACM with the keyword 'user experience management' and variations of it. In the end, we found some relevant papers. In these papers, there are various approaches or descriptions of UX management. The term UX management is often used without any explanation. We present these papers in the next paragraph.

A. UX Management

The term UX management is used differently in literature. The main task of a UX manager, according to Szóstek [9], is the development of the UX team. Szóstek [9] describes the development of the team with the selection of the best career path for individual team members. This includes career planning and development, team management, and training of individual team members. UX management in this case is related to team building and empowerment.

In addition to building a UX team, Anderson et al. [10] proposed that C-level executives should be involved in it. C-level executives should understand that UX management is necessary to develop products with a high user experience. With the cooperation of C-level executives, UX teams can work successfully. For the implementation of UX management, Anderson et al. [10] and Rosenberg [11], for example, offer various patterns that provide support at the levels of planning, decision, tactics, and conflict.

B. UX Maturity Models

The use of UX Maturity Models is one way to at least measure the current state of implementation of UX activities within an organization. The advantage of using such a model is that it determines the current maturity level of the organization. Thus, its weaknesses can be identified. The result can be used to work specifically on improving UX Maturity. There are different UX Maturity Models that measure various aspects.

The *Total User Experience Management* (*TUXM*) [12] model contains elements such as UX objectives, integrated design system, strategic communication, continual improvement, fact-based decision-making, and a T-type design team. The N*ielsen Corporate Usability Maturity Model* [13], on the other hand, comprises dimensions such as the developers' attitude towards usability, the management's attitude towards usability, the usability practitioner's role, usability methods and techniques, and strategic usability. Another approach is the metric *Index of Integration* (*IoI*) [14]. This metric can be used to determine the maturity level of typical HCI activities in the development team.

It is noticeable that the approaches presented capture different dimensions of UX management. For example, the TUXM model measures the dimension 'UX objectives', which is not present in *Nielsen Corporate Usability Maturity Model.* The metric *Index of Integration* (*IoI*) in turn only includes HCI activities. Conversely, the Nielsen model is more focused on practical implementation. The testing of a suitable UX maturity model should be carried out before deployment and tailored to the needs of the organization [13].

C. UX Methods in Agile Development

In the literature, various UX methods are used in agile development. In the study of Hinderks et al. [15], 16 UX methods used in agile development were identified. The two most frequently used methods are *Prototyping* and *Personas*. Prototyping and personas can be used as artifacts for the communication between UI designers and developers. The UI designers either develop a prototype together with the developers, or work on it before the actual development. Personas, on the other hand, are usually used permanently. Various methods are used to determine the requirements— these are ta*sk/usage scenarios, focus groups, contextual inquiry, user evaluation, interviews, A/B testing, card sorting, brainstorming,* and *FlexREQ*. The following methods are used to measure and evaluate the user experience: acceptance test, expert reviews, UX questionnaires, usability testing, and usability inspection. It was impossible to determine at what stage (before, during, or after development) the UX methods were used.

D. GAP Analysis

We generally understand management based on the explanations of Drucker [16] and Stone [17]—it is a combination of a goal, a strategy, and resources.

When applied to UX, user experience management consists of a UX goal, a UX strategy, and UX resources (Fig. 1) based on the work of McKeown [18].

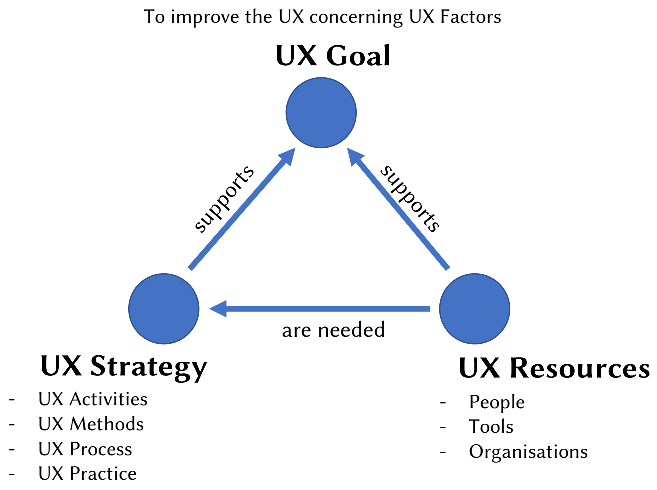


Fig. 1. User Experience Management based on McKeown [18].

For example, a UX goal can be set upfront based on user research to improve the UX for a selected factor of the UX. This can be, for example, the UX factor 'Trust'. To improve the UX factor 'Trust', a UX strategy can be developed by using various UX methods, which is then implemented by a UX team (UX resources). A subsequent evaluation with the User Experience Questionnaire Plus (UEQ+) [19] or the SUPR-Q [20] can be performed. In addition, a benchmark [21] or KPI [22] can be calculated based on the individual UX factors. The UEQ+ is a modular framework that allows one to combine predefined UX factors to create a concrete UX questionnaire. Currently, the UEQ+ framework contains 20 UX scales, but they can be extended as needed. The construction of the clarity factor can be read as an example [23]. The result can be used to determine whether the previously defined UX goal has been achieved or not.

Both UX strategy and UX resources are necessary to achieve the UX goal. It should be known before the next development iteration, whose requirements positively supported the UX goal. In this way the UX goal can be achieved in a goal-oriented manner.

In our view, it makes absolute sense to empower and develop a UX team. This is a necessary prerequisite to be able to successfully implement UX management at all. In our opinion, however, a UX goal and a UX strategy are also needed to be able to operate UX management successfully.

For this reason, we conducted a tertiary study to identify requirements from the analysed literature reviews that had their focus on user experience and agile development.

III. Research Methodology

In order to answer our research question, the first step is to conduct a literature review. We conducted the study according to the guidelines for SLR in software engineering by Kitchenham and Charters [24].

We used two main tools to conduct the study. We used the SLR tool [25] for conducting the search (managing the paper, developing the review protocol, documenting the search, and conducting quality assessment). In our literature database managed with Citavi, we imported the result of the SLR from the SLR tool to use the management and citation functions.

A. Objectives and Research Question

However, during the research on the related work (Section II), we also found that the term 'UX management' is neither sufficiently defined nor explained in the literature. Further, we found through GAP analysis (Section II.D) that there was a research gap in the goal, strategy, and resources concerning UX management.

In this context, we would like to answer the following research question:

RQ: What requirements can be derived from literature reviews for User Experience Management with the focus on agile development?

This question aims to identify requirements from the literature that can potentially be adopted for UX management in agile development. Our goal is then to create a consolidated list of requirements for UX management. This can then be used in practical implementation. Also, UX activities or UX processes can be derived based on this list of requirements.

B. Search Strategy and Data Sources

Based on the research question we have developed a search strategy. This strategy consists of a search string, the search space, and the process to select relevant papers.

Our search string consists of three groups, each covering one area. These are 'agile', 'user experience', and 'literature reviews'. Both 'agile' and 'user experience' are necessary search terms to narrow down the set of topics. We further decided to conduct a tertiary study. Accordingly, we extended the search string to 'literature reviews', since we wanted to base our study on literature reviews that had already been conducted.

In a second step, we collected possible keywords for each group of the search string and extended them with alternative spellings and synonyms. The search string developed in this way is as follows:

(agile OR kanban OR scrum OR lean OR "extreme
programming" OR "design thinking")
AND
("user experience" OR ux OR usability OR hcd OR hci OR hmi OR ucd)
AND
("SLR" OR "Structured literature reviews" OR "mapping study" OR
"systematic review")

This search string was logically adapted to the syntax of the search spaces. The search space included digital libraries, journals, and conference proceedings. A complete list of the search space is shown in Table I.

TABLE I. Search Space With Specification of Search Strategy (TAK = Title, Abstract, and Keywords) and Number of Papers

Library	Search Strategy	Number
SpringerLink	Full Text	974
IEEE Xplore	Full Text	8
SCOPUS	TAK	41
Science Direct	TAK	1
ACM	TAK	1

The search was conducted at all search spaces in February 2021. Without any restriction, that is plain full-text search of the search engine, $N_{P0} = 4{,}363$ papers were found.

C. Study Selection

The results from the individual search spaces were imported into the SLR tool. Duplicate entries were already removed during the import. As a result, 1,023 papers were imported and then analysed in further steps as described in the next paragraphs. The result is shown in Table II.

TABLE II. Search Process Comprising Phases

ID	Method	Base	Reduced	Res.
N_{P1}	Extended search	1,023	-977	46
N_{P2}	After scan title	46	-22	24
N_{P3}	After scan abstract	24	-2	22
N_{P4}	Apply quality criteria	22	-5	16
N_{P5}	**Final dataset**	**16**		

By using the internal search function of the SLR Tool [25], we were able to reduce the result by searching only on title, abstract, and keywords from $N_{P1} = 1{,}023$ to $N_{P2} = 46$. This was necessary because the search space 'IEEE Xplore' and 'SpringerLink' were explored initially through a full text-search.

In a further step, we reduced the number of papers from $N_{P2} = 46$ to $N_{P3} = 24$ going by the title. We only included those papers that were interesting and valuable for our study in terms of title. One should be able to readily recognize from its title that the paper is mainly about 'agile' and 'user experience'. Additionally, it should be a literature review. In the following step, we reduced the number of papers to $N_{P4} = 22$ going by the abstract. We applied the same criteria we used one step ago. All the decisions were traceably logged by the SLR tool.

The papers selected in the step before ($N_{P4} = 22$) were evaluated with a quality assessment. In the assessment, the papers were checked to see whether the literature review was carried out in a traceable manner. Also, we checked whether the literature review was performed according to a standard published in the literature. The papers were then reduced to $N_{P5} = 16$.

In each step of the reduction, a set of selection criteria were applied. They were then divided into inclusion and exclusion criteria. The inclusion criteria were: papers written in English; peer-reviewed papers; and papers presenting literature review to integrate user experience methods (or similar) into agile development processes. Exclusion criteria were: papers whose full text were not available; papers with results that had already been published; and papers that were not focused on agile development.

IV. Results

In our work, we have selected 16 relevant studies. The first part of this section gives an overview of the selected studies. In the second part, the individual research questions will be answered based on the studies.

TABLE III. Overview of the Included SLRs in Our Tertiary Study

ID	Author	Title	
[26]	Bruun	Training software developers in usability engineering	2010
[27]	Silva et al.	User-Centered Design and Agile Methods: A Systematic Review	2011
[28]	Salvador et al.	A systematic review of usability techniques in agile methodologies	2014
[29]	Jurca et al.	Integrating Agile and User-Centered Design: A Systematic Mapping and Review of Evaluation and Validation Studies of Agile-UX	2014
[30]	Salah et al.	A systematic literature review for agile development processes and user centred design integration	2014
[31]	Silva et al.	A Systematic Mapping on Agile UCD Across the Major Agile and HCI Conferences	2015
[32]	Brhel et al.	Exploring principles of user-centered agile software development: A literature review	2015
[33]	Magues et al.	Usability in agile development: A systematic mapping study	2016
[34]	Magues et al.	HCI usability techniques in agile development	2016
[35]	Caballero et al.	How Agile Developers Integrate User-Centered Design Into Their Processes: A Literature Review	2016
[36]	Garcia et al.	Artifacts for Agile User-Centered Design: A Systematic Mapping	2017
[37]	Hoda et al.	Systematic literature reviews in agile software development: A tertiary study	2017
[38]	Schön et al.	Agile Requirements Engineering: A systematic literature review	2017
[39]	Pereira et al.	Design Thinking Integrated in Agile Software Development: A Systematic Literature Review	2018
[40]	Silva et al.	The evolution of agile UXD	2018
[41]	Curcio et al.	Usability in agile software development: A tertiary study	2019

A. Overview of the Studies

Our search was limited to 'Agile', 'UX', and 'SLR'. An explicit restriction to UX mangement or similar was not made. Our experience in an initial literature search was that restricting by 'UX management' did not yield useful results. For this reason, we found publications dealing with the integration of 'Agile' and 'UX'. The search result then served us as a basis for deriving requirements for UX management.

In Table III we listed all included SLRs under the terms of author, title, and year.

In a further step, we investigated the research questions of SLRs. We wanted to figure out which of the research questions UX strategy, UX ressources, or UX goal was addressed. The classification was made based only on the purposes of the research questions. A total of 47 research questions from the 16 SLRs were examined. Twenty-nine research questions were assigned to the category UX strategy, 7 to UX resources, and 0 to UX goal. The remaining 11 research questions could not be assigned to any of the categories.

B. RQ: What Requirements Can Be Derived From Literature Reviews With the Focus on User Experience and Agile Development?

To answer our research questions, we examined each SLR in terms of the core statement and a corresponding categorization. We presented the result in Table IV. In doing so, we extracted the core statement as a citation from the SLR if it was possible. When this was not possible, we created our own summary.

We categorized the statements concerning the type of investigation. This classification will help us later to derive the requirements from the statements. We distinguish as follows:

- **Finding**: For us, a finding is an insight into agile development in terms of user experience. The insight or statement can be positive or negative. It can also be a recommendation based on the results found.
- **Problem**: Compared to a finding, the naming of a problem is more concrete and specific. This means that a problem is a substantial challenge.
- **Method**: The category method includes methods recommended or used in the area of agile development and user experience.

In the next sections, we present the findings, problems, and methods identified.

1. Findings

Hoda et al. [37] state that the integration of user experience and agile software development has not made significant progress between the years 1990 and 2015. According to the authors, there is still the challenge to combine research rigor with industrial relevance.

The lack of continuous involvement of stakeholders [32], especially in requirements elicitation and analysis [33] has been identified as a problem space. Failure to involve stakeholders in the early stages of requirements elicitation results in such a failure to create a shared understanding of the product and its goal [38]. As a result, the user perspective in agile software development (ASD) is not well established [38].

In addition to continuous involvement of stakeholders, Brhel et al. [32] derived four additional principles: separate product discovery and product creation, iterative and incremental design and development, parallel interwoven creation tracks, and artefact-mediated communication.

Silva da Silva et al. [31] answered the research questions on how Agile UCD was understood and which techniques were used in Agile UCD. In addition, the authors identified the benefits that could be gained from integrating agile software development and user experience. The benefits were as follows: improved communication, improved visibility, customer input during the release, business analysis improvement, priorization of the backlog, and improved usability.

TABLE IV. Overview Statements From the SLRs

Year	Author	Statement	Category
2010	Bruun [26]	"Fifth, as discussed previously, user based evaluation methods seem to provide the best wake-up call for software companies..."	Method
2011	Silva et al. [27]	"A very important point is to maintain the Big Picture, which is difficult given the characteristic of iterative development in agile projects."	Problem
		"The focus of integrating agile methods and UCD should be on design as well as on usability evaluation."	Method
2014	Salvador et al. [28]	"The most commonly used usability methods are: fast prototyping, individual inquiry, formal tests, and heuristic evaluations."	Method
2014	Jurca et al. [29]	"One issue that was common between most validation and evaluation studies, was the power struggle between UX designers and developers."	Problem
		"Furthermore, there are often not enough UX designers involved in the Agile projects."	Problem
2014	Salah et al. [30]	The identified key aspects are: lack of allocated time for upfront activities, difficulty of modularization, optimizing the work between developers and UCD practitioners, performing usability testing, and lack of documentation.	Problem
2015	Brhel et al. [32]	"The analysis resulted in a comprehensive coding system and five principles for UCASD: (1) separate product discovery and product creation, (2) iterative and incremental design and development, (3) parallel interwoven creation tracks, (4) continuous stakeholder involvement, and (5) artifact-mediated communication."	Finding
2015	Silva et al. [31]	The most commonly used HCI techniques are Usability testing on lightweight prototypes, Continuous Research, Evolutionary Prototyping, Upfront Design, Personas.	Method
		The benefits in integrating agile and UX are improved communication, improve visibility, customer input during the release, business analysis improvement, priorization of the backlog, and improved usability.	Finding
2016	Magues et al. [33]	"In conclusion, the authors concluded that there are no formalised suggestions for integrating usability techniques into agile software development."	Problem
2016	Magues et al. [34]	"Most of the human-computer interaction (HCI) techniques that the ASDP is adopting are techniques related to requirements engineering, especially techniques for requirements elicitation and analysis."	Finding
2016	Caballero [35]	The three main UCD methods, which represent 70% of the methods used, are Prototypes, User Stories, and Usability Testing.	Method
2017	Schön et al. [38]	"Based on a qualitative analysis of the included studies, we can conclude that building a shared understanding of the user perspective is not very well established in ASD."	Finding
		"We identified the following key artifacts for the documentation of requirements that are used in Agile RE: User stories, prototypes, use cases, scenarios and story cards."	Method
2017	Hoda [37]	"We did not find much evidence to support a significant progress toward resolving the 'grand challenge' of ASD: combining research rigor with industrial relevance, as a topic."	Finding
		The four most commonly used artifact used to faciliate communication are prototype, user story, cards, and persona.	Method
2018	Silva et al. [40]	"Finally, the authors concludes that Technology and Artifacts are still missing to achieve integration between Agile Methods and User Experience Design to Agile UXD."	Problem
2019	Curcio et al. [41]	"Regarding to the challenges for the integration seven main categories were also identified: issues related to tests, time, work balance, modularization, feedback, prioritization, and documentation."	Problem

2. Problems

One problem Silva da Silva addresses relates to the development and sustenance of a Big Picture [27]. This is openly a problem of iterative development. Because of the fact that requirements can change from iteration to iteration, the Big Picture can also change.

Another problem is the available capacity of UX professionals. Jurca et al. [29] found that there was often not enough capacity of UX professionals for individual projects. Another finding the authors noted was that the relationship between UX designers and developers tended to be poor, which negatively impacted the outcome of the collaboration. This conclusion was also reached by Salah et al. [30]. Optimizing the work between UX professionals and developers is a key aspect here [30]. In general, it seems that there is too little capacity for UX activities [30], [41].

According to Magües et al. [33], the integration of usability techniques into agile development is not formalized or institutionalized. This means that there are no formalized proposals for the integration of usability techniques or methods into agile development. As a result, the use of usability techniques may not deliver the desired result and, thus, the actual potential is lost.

3. Methods

Various artifacts have been established for communication and documentation. These are, for example, user stories, prototypes, use cases, story cards, and personas [31], [35], [36], [38]. The use of the artifacts not only facilitates communication between stakeholders and the development team [38], but also, as in the case of the prototype, provides the basis for an evaluation of the result [27].

It is important to note that UX or usability evaluation is one of the most frequently mentioned methods in SLRs [26]–[28], [31], [35]. Specifically, the evaluation of a prototype in the early stages of development is indicated as the most commonly used method. The goal is to gain new knowledge as early as possible, which can then be incorporated back into the development. The evaluation itself is carried out in a variety of ways which are not mentioned here.

Other UX or usability methods that can be named are individual inquiry, formal tests, heuristic evaluations [28], continuous research, upfront design, and personas [31].

4. Summary of Requirements

From the results, requirements for UX management can be

People/Social

- Improve and optimize collaboration between UX professionals and developers [25], [26], [27], [37].
- Create sufficient capacity of UX designers [25].
- Continuous stakeholder involvement [27], [28], [30], [34].

Technology/Artifact

- Apply evaluation methods such as usability evaluation, formal test, heuristics evaluation [22], [24], [26], [27], [31].
- Create and evaluate prototypes [24], [27], [31], [32].
- Use of artifacts such as user stories, cards, and personas to communicate artefact-mediated [28], [31], [32], [34].
- Create documentation [26], [34].

Process/Practice

- Create a Big Picture of the product [27].
- Formalised suggestions for integrating usability techniques into agile software development [29].
- Separate product discovergy and product creation [27], [28].
- Parallel interwoven creation tracks [28].
- Iterative and increment research, design, and development [27], [28].
- Allocated time for upfront activities [26].

Fig. 2. Overview of Requirements Categorised by People/Social, Technology/Artifacts, and Process/Practice.

formulated in summary. This means what should be implemented and how, or which method should be used. The following list in Fig. 2 is based on the summaries in Table IV.

We have classified the requirments according to Brhel et al. [32]. The three categories are People/Social, Technology/Artifact, and Process/Practice. To us, the division seems reasonable because the requisitions address different areas in the implementation.

As can be seen from the distribution of individual requirements, some requirements have been addressed in multiple SLRs, such as prototypes or evaluation methods. In the next section, we discuss the results in greater detail.

V. Discussion

Notwithstanding, it is clear from the requirements that prototypes and UX evaluation are established methods. The use of prototypes, in whatever form, was indicated as most commonly used methods in five of sixteen SLRs. UX evaluation methods were indicated in six out of sixteen SLRs. These two methods should definitely be part of UX management.

In this context, the question also arises as to when different UX methods should or can be used. In Hinderks [15], 18 approaches were analysed with regard to their temporal applicability in development. The aim was to examine the phase of development for each approach, in which the approach was to be, or was, used. The breakdown was structured according to whether the approach had been applied before, during, or after development. If an approach could be used in several phases of development, it was also assigned to those phases. In total, 21 (88%) approaches can be used before development while 15 (63%) approaches can be used during development, and 13 (54%) after development.

A. People/Social

It is not clear from the requirements whether there is one team responsible for product development and UX management, or whether there is an additional UX team that handles UX activities. In the first case, the UX professionals would be members of the team and would perform the UX activities during an iteration. In the second case, there is a UX team that handles the UX activities for multiple product teams. Based on the problems described [29], [30], obviously, there are still knowledge deficits as to how UX professionals and developers can work together smoothly or without problems. Furthermore, it cannot be determined whether the described problems arise due to the fact that the UX professionals are part of the product team or work for the product team. From the findings (Fig. 2), it can be concluded that UX management should improve and optimize the collaboration between UX professionals and developers. In addition, capacity for UX designers should be created. The requirement also matches the results from Hinderks [42]. UX Poker is a method to estimate the UX expected for a requirement per UX factor. The method has been conducted with UX professionals and developers. It has been shown that a common understanding about the UX of the product to be developed could be gained.

Stakeholders should also be integrated into the processes as a further requirement. This is also the requirement of Human-Centred Design (HCD) [2], for example. Human-Centred Design (HCD) is an approach to develop user-centred products by putting the user at the centre of the development process. The idea behind HCD is to develop a great understanding of the user and their requirements. The focus is placed on the user through the iterative process and continuous testing of alternative solutions. HCD itself does not describe the collaboration between UX professionals and developers. In this respect, HCD can only be a partial solution and can only be used in combination with other methods or processes.

B. Technology/Artefact

UX evaluation can be named as one of the most frequently mentioned requirements for UX management. Different methods can be used for evaluation. In addition to the developed product, prototypes should also be evaluated. Regarding UX management, we are convinced that the evaluation of prototypes as well as the developed product should play a decisive role in it. Only when evaluating, can it be determined whether the UX goal has been reached at all. It always makes sense to do a UX evaluation, but with a goal it is more focused. And with a UX goal, UX management can be performed. Another requirement is the use of different artefacts, such as user stories, cards, and personas. The use of artefacts supports communication between stakeholders, UX professionals, and developers.

C. Process/Practice

In the category Process/Practice, there are partly very specific approaches, such as *create a big picture of the product or allocated time for upfront activities*. On the other hand, general approaches are also mentioned under it, such as *parallel interwoven creation tracks, allocated time for upfront activities*, or *formalized suggestions for integrating usability techniques into agile software development*. In the sense of a process, *separate product discovery and product creation and iterative and increment research, design, and development* can be classified. How and when the corresponding processes or practices are to be used is not clear from the literature. It is also not clear how they can be used in relation to UX management.

An important conclusion from Section IV is that UX evaluation is an important method to apply. In this regard, the question is who performs the UX evaluations and when. From the results, it can be determined that UX evaluations are performed both before, during and after development. However, it is not suggested who should perform these evaluations. Classically, it is the responsibility of UX Researchers who are either integrated into the team or work for it. Also, ResearchOps [43], [44] can facilitate the required foundation for a UX Research by offering roles, tools and processes.

D. UX Goal

No UX goal can be directly derived from the results found. The analysis of the research questions of the SLR also reveals that none of these can be assigned to a UX goal. Twenty-nine research questions have been mapped to UX strategy, which is largely reflected in the requirements 4. This is remarkable because a UX strategy and UX resources are supposed to support the achievement of a UX goal. However, if no UX goal has been named, the success of the UX strategy cannot really be measured. While every UX strategy also has an impact, in our opinion this should be managed from goals.

E. UX Management

From the results presented in Section IV and requirements listed in Fig. 2, the requirements for UX management can be summarized into two key requisites:

- Enable and support collaboration between stakeholders, UX professionals, and developers.
- Evaluate the user experience in the context of a UX goal.

All other UX methods or approaches found can be assigned to one of the key requirements. The specific UX methods used are interchangeable or replaceable with other methods. For example, requirements can be collected as a user story. A user story can be used to capture a requirement briefly and comprehensibly. In addition, the artefact user story is very well suited for collaboration between stakeholders and UX professionals. Another example, questionnaires can be used for the evaluation of UX. But other evaluation methods can also be used. All that matters is that the UX is evaluated and the result can be compared to a UX goal. It should be decided in the specific project or team which methods are to be used.

F. Limitations

We have collected the requirements for UX management based on a literature review. Further studies on UX professionals, developers, and managers should validate or complement the list of requirements for UX management created in this study.

VI. Conclusion and Future Work

This paper presents a tertiary study of UX management to identify potential requirements for user experience management. The tertiary study was conducted according to the guideline offered by Kitchenham and Charters [24]. In an initial search, we found 4,363 studies. Our search process reduced the number of studies to 1,023. We analysed these studies by their titles and abstracts and performed a quality assessment. Finally, we selected and further analysed 16 studies.

The requirements we identified all related to UX methods or improvements in the software development process. The two most frequently mentioned UX methods were prototyping and UX or usability evaluation. Communication between UX professionals and developers was identified as a major improvement in the software development process.

We also analysed the research questions of the SLRs with regard to a possible assignment to UX goal, UX strategy, and UX resourcces. We wanted to determine to what extent all three areas were covered. For UX strategy and UX resourcces, we were able to identify corresponding research questions. For UX goal, we could not find any research questions or requirements.

In summary, we were able to identify requirements for UX management. However, we could not identify requirements for UX management that enabled the development and achievement of a UX goal.

References

[1] J. Preece, Y. Rogers, H. Sharp, *Interaction design: Beyond human-computer interaction.* Chichester: Wiley, 4. d. ed., 2015.

[2] ISO9241-210, “Ergonomics of human-system interaction - part 210: Human-centred design for interactive systems,” 2020.

[3] K. Schwaber, J. Sutherland, *The Scrum Guide: The Definitive Guide to Scrum: The Rule of the Game.* 2020.

[4] K. Schwaber, *Agile project management with Scrum.* Microsoft professional, Redmond, Wash.: Microsoft Press, 2004.

[5] D. J. Anderson, *Kanban: Successful evolutionary change for your technology business.* Sequim, Washington: Blue Hole Press, 2010.

[6] K. Beck, C. Andres, *Extreme programming explained: Embrace change.* The XP series, Boston: Addison- Wesley, 2. ed., 6. printing ed., 2007.

[7] P. Serrador, J. K. Pinto, “Does agile work? — a quantitative analysis of agile project success,” *International Journal of Project Management*, vol. 33, no. 5, pp. 1040–1051, 2015, doi: 10.1016/j.ijproman.2015.01.006.

[8] B. Boehm, R. Turner, “Using risk to balance agile and plan- driven methods,” *Computer*, vol. 36, no. 6, pp. 57– 66, 2003, doi: 10.1109/MC.2003.1204376.

[9] A. Szóstek, “A look into some practices behind microsoft ux management,”in *Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts - CHI EA ’12*, New York, New York, USA, 2012, p. 605, ACM Press.

[10] R. I. Anderson, J. Ashley, T. Herrmann, J. Miller, J. Nieters, S. S. Eves, S. T. Watson, “Moving ux into a position of corporate influence,” in *CHI ’07 extended abstracts on Human factors in computing systems - CHI ’07*, New York, New York, USA, 2007, p. 1905, ACM Press.

[11] D. Rosenberg, “The business of ux management,” *interactions*, vol. 26, no. 3, pp. 28–35, 2019, doi: 10.1145/3318131.

[12] H. B.-L. Duh, J.-J. Lee, P. L. P. Rau, M. Q. Chen, “The management model development of user experience design in organization,” in *Cross-Cultural Design*, vol. 9741 of *Lecture Notes in Computer Science*, P.-L. P. Rau Ed., Cham: Springer International Publishing, 2016, pp. 163–172, doi: 10.1007/978-3-319-40093-8_17.

[13] D. Salah, R. Paige, P. Cairns, “Integrating agile development processes and user centred design- a place for usability maturity models?,” in *Human-Centered Software Engineering*, vol. 8742 of *Lecture Notes in Computer Science*, S. Sauer, C. Bogdan, P. Forbrig, R. Bernhaupt, M. Winckler Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 108–125, doi: 10.1007/978-3-662-44811-3_7.

[14] A. Joshi, N. L. Sarda, S. Tripathi, “Measuring effectiveness of hci integration in software development processes,” *Journal of Systems and Software*, vol. 83, no. 11, pp. 2045–2058, 2010, doi: 10.1016/j.jss.2010.03.078.

[15] A. Hinderks, F. J. Domínguez Mayo, J. Thomaschewski, M. J. Escalona, “Approaches to manage the user experience process in agile software development: A systematic literature review,” *Information and Software Technology*, vol. 150, p. 106957, 2022, doi: 10.1016/j.infsof.2022.106957.

[16] P. F. Drucker, *The practice of management.* New York, NY: HarperCollins, 2009.

[17] J. Magretta, N. D. Stone, *What management is: How it works and why it’s everyone’s business.* London: Profile Books, 2013.

[18] M. Mckeown, *The strategy book: How to think and act strategically to deliver outstanding results.* Harlow, England: Pearson, 3rd edition ed., 2020.

[19] 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, 2019, doi: 10.9781/ijimai.2019.06.006.

[20] J. Sauro, “Supr-q: A comprehensive measure of the quality of the website user experience,” *Journal of Usability Studies*, vol. 2015, no. 10, pp. 68–86, 2015.

[21] M. Schrepp, A. Hinderks, J. Thomaschewski, “Construction of a benchmark for the user experience questionnaire (ueq),” *International Journal of Interactive Multimedia and Artificial Inteligence*, vol. 4, no. 4, pp. 40–44, 2017, doi: 10.9781/ijimai.2017.445.

[22] A. Hinderks, M. Schrepp, F. J. Domínguez Mayo, M. J. Escalona, J. Thomaschewski, “Developing a ux kpi based on the user experience questionnaire,” *Computer Standards & Interfaces*, vol. 65, pp. 38–44, 2019, doi: 10.1016/j.csi.2019.01.007.

[23] M. Schrepp, R. Otten, K. Blum, J. Thomaschewski, “What causes the dependency between perceived aesthetics and perceived usability?,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 2020, no. 6, pp. 78–85, 2020, doi: 10.9781/ijimai.2020.12.005.

[24] B. Kitchenham, S. Charters, “Guidelines for performing systematic literature reviews in software engineering,” 2007.

[25] A. Hinderks, M. Schrepp, F. J. Domínguez Mayo, M. J. Escalona, J. Thomaschewski, “Ueq kpi value range based on the ueq benchmark.” doi: 10.13140/RG.2.2.34239.76967.

[26] A. Bruun, “Training software developers in usability engineering,” in *Proceedings of the 6th Nordic Conference on Human-Computer Interaction Extending Boundaries - NordiCHI ’10*, New York, New York, USA, 2010, p. 82, ACM Press.

[27] T. Silva da Silva, A. Martin, F. Maurer, M. Silveira, “User-centered design and agile methods: A systematic review,” in *2011 AGILE Conference*, 2011, pp. 77–86, IEEE.

[28] C. Salvador, A. Nakasone, J. A. Pow-Sang, “A systematic review of usability techniques in agile methodologies,” in *Proceedings of the 7th Euro American Conference on Telematics and Information Systems - EATIS ’14*, New York, New York, USA, 2014, pp. 1–6, ACM Press.

[29] G. Jurca, T. D. Hellmann, F. Maurer, “Integrating agile and user-centered design: A systematic mapping and review of evaluation and validation studies of agile- ux,” in *2014 Agile Conference*, 2014, pp. 24–32, IEEE.

[30] D. Salah, R. F. Paige, P. Cairns, “A systematic literature review for agile development processes and user centred design integration,” in *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE ’14*, New York, New York, USA, 2014, pp. 1–10, ACM Press.

[31] T. Silva da Silva, F. F. Silveira, M. S. Silveira, T. Hellmann, F. Maurer, “A systematic mapping on agile ucd across the major agile and hci conferences,” in *Computational Science and Its Applications – ICCSA 2015*, vol. 9159 of *Lecture Notes in Computer Science*, O. Gervasi, B. Murgante, S. Misra, M. L. Gavrilova, A. M. A. C. Rocha, C. Torre, D. Taniar, B. O. Apduhan Eds., Cham: Springer International Publishing, 2015, pp. 86–100, doi: 10.1007/978-3-319-21413-9_7.

[32] M. Brhel, H. Meth, A. Maedche, K. Werder, “Exploring principles of user-centered agile software development: A literature review,” *Information and Software Technology*, vol. 61, pp. 163–181, 2015, doi: 10.1016/j.infsof.2015.01.004.

[33] D. A. Magües, J. W. Castro, S. T. Acuña, “Usability in agile development: A systematic mapping study,” in *2016 XLII Latin American Computing Conference (CLEI)*, 2016, pp. 1–8, IEEE.

[34] D. A. Magues, J. W. Castro, S. T. Acuna, “Hci usability techniques in agile development,” in *2016 IEEE International Conference on Automatica (ICA-ACCA)*, 2016, pp. 1–7, IEEE.

[35] L. Caballero, A. M. Moreno, A. Seffah, “How agile developers integrate user-centered design into their processes: A literature review,” *International Journal of Software Engineering and Knowledge Engineering*, vol. 26, no. 08, pp. 1175–1201, 2016, doi: 10.1142/S0218194016500418.

[36] A. Garcia, T. Silva da Silva, M. Selbach Silveira, “Artifacts for agile user-centered design: A systematic mapping,” in *Proceedings of the 50th Hawaii International Conference on System Sciences (2017)*, Proceedings of the Annual Hawaii International Conference on System Sciences, 2017, Hawaii International Conference on System Sciences.

[37] R. Hoda, N. Salleh, J. Grundy, H. M. Tee, “Systematic literature reviews in agile software development: A tertiary study,” *Information and Software Technology*, vol. 85, pp. 60–70, 2017, doi: 10.1016/j.infsof.2017.01.007.

[38] E.-M. Schön, J. Thomaschewski, M. J. Escalona, “Agile requirements engineering: A systematic literature review,” *Computer Standards & Interfaces*, vol. 49, pp. 79– 91, 2017, doi: 10.1016/j.csi.2016.08.011.

[39] J. C. Pereira, R. d. F. Russo, “Design thinking integrated in agile software development: A systematic literature review,” *Procedia Computer Science*, vol. 138, pp. 775– 782, 2018, doi: 10.1016/j.procs.2018.10.101.

[40] T. S. Da Silva, M. S. Silveira, F. Maurer, F. F. Silveira, “The evolution of agile uxd,” *Information and Software Technology*, vol. 102, pp. 1–5, 2018, doi: 10.1016/j.infsof.2018.04.008.

[41] K. Curcio, R. Santana, S. Reinehr, A. Malucelli, “Usability in agile software development: A tertiary study,” *Computer Standards & Interfaces*, vol. 64, pp. 61– 77, 2019, doi: 10.1016/j.csi.2018.12.003.

[42] A. Hinderks, D. Winter, F. J. Domínguez Mayo, M. J. Escalona, J. Thomaschewski, “ Ux poker: Estimating the influence of user stories on user experience in early stage of agile development,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 7, pp. 97-104, 2022, doi: 10.9781/ijimai.2022.11.007.

[43] M. de Bayser, L. G. Azevedo, R. Cerqueira, “Researchops: The case for devops in scientific applications,” in *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, 2015, pp. 1398–1404, IEEE.

[44] E. Savarit, “Fitting user research into your organization,” in *Practical User Research*, E. Savarit Ed., Berkeley, CA: Apress, 2020, pp. 47–79, doi: 10.1007/978-1-4842-5596-4_3.

Andreas Hinderks

Andreas Hinderks holds a PhD in Computer Science by University of Sevilla. He has worked in various management roles as a Business Analyst and a programmer from 2001 to 2016. His focus lay on developing user-friendly business software. Currently, he is a freelancing Product Owner, Business Analyst and Senior UX Architect. He is involved in research activities dealing with UX questionnaires, measuring user experience and User Experience Management since 2011.

Francisco José Domínguez Mayo

Francisco José Domínguez Mayo received a PhD degree in computer science from the University of Seville, Seville, Spain, in July 2013. He is currently an associate professor with the Department of Computing Languages and Systems, University of Seville. He collaborates with public and private companies in software development quality and quality assurance. The focus of his interesting research is on the areas of continuous quality improvement and quality assurance on software products, and software development processes.

María José Escalona

María José Escalona received her PhD in Computer Science from the University of Seville, Spain in 2004. Currently, she is a Full Professor in the Department of Computer Languages and Systems at the University of Seville. She manages the web engineering and early testing research group. Her current research interests include the areas of requirement engineering, web system development, model-driven engineering, early testing and quality assurance. She also collaborates with public companies like the Andalusian Regional Ministry of Culture and Andalusian Health Service in quality assurance issues.

Jörg Thomaschewski

Jörg Thomaschewski received a PhD in physics from the University of Bremen (Germany) in 1996. He became a Full Professor at the University of Applied Sciences Emden/Leer (Germany) in September 2000. His teaching and research focus is on Human-Computer Interaction, UX-Management, Agile Software Development, and Requirements Engineering. Since 2012 he has been the head of the research group ‘Agile Software Development and User Experience’. Dr. Thomaschewski has extensive experience in user experience training, UX questionnaires, agile methods, IT analysis, and consulting.

Longitudinal Segmented Analysis of Internet Usage and Well-Being Among Older Adults

Alejandro Cervantes[1], David Quintana[2]*, Yago Sáez[2], Pedro Isasi[2]

[1] Escuela Superior de Ingeniería y Tecnología, Universidad Internacional de La Rioja (UNIR), Logroño (Spain)

[2] Computer Science and Engineering Department, Universidad Carlos III de Madrid, Leganés (Spain)

* Corresponding author. dquintan@inf.uc3m.es

Received 8 June 2020 | Accepted 3 March 2022 | Early Access 2 May 2022

Abstract

The connection between digital literacy and the three core dimensions of psychological well-being is not yet well understood, and the evidence is controversial. We analyzed a sample of 2,314 individuals, aged 50 years and older, that participated in the English Longitudinal Study of Aging. Participants were clustered according to drivers of psychological well-being using Self-Organizing Maps. The resulting groups were subsequently studied separately using generalized estimating equations fitted on 2-year lagged repeated measures using three scales to capture the dimensions of well-being and Markov models. The clustering analysis suggested the existence of four different groups of participants. Statistical models found differences in the connection between internet use and psychological well-being depending on the group. The Markov models showed a clear association between internet use and the potential for transition among groups of the population characterized, among other things, by higher levels of psychological well-being.

Keywords

Aging, ELSA, Internet, Markov Models, Psychological Wellbeing, Self-Organizing Maps.

DOI: 10.9781/ijimai.2022.05.002

I. Introduction

The connection between psychological well-being and digital literacy at advanced age is an open research question at the core of a growing number of studies. Among them, only a few rely on large samples that track participants over long periods of time.

The aim of this study is providing further insights on the connection between Internet use and psychological well-being at advanced age using well-known artificial intelligence methods. The main contribution will be testing whether modeling the population as a homogeneous set causes a loss of relevant information that might be revealed by a more fine-grained segmented analysis.

We suggest clustering the population using machine-learning to subsequently fit more traditional statistical models on specific segments to assess the differential impact, if any, of Internet use on three core aspects of psychological well-being. This poses an innovation in this context that could potentially help identify connections that might have been overlooked in the literature.

We also intend to enrich the analysis exploring whether digital literacy results in differences in transition dynamics among the identified clusters over time using Markov models. This is relevant because it might show interesting patterns regarding the transition from clusters associated with higher degrees of psychological well-being to the ones with the lower one, and vice-versa.

The idea of the initial clustering, the dynamic analysis based on cluster transition, and the use of these algorithms all represent, to the best of our knowledge, technical innovations in the study of psychological well-being and digital literacy at advanced age.

The rest of the paper is organized as follows: section II will be devoted to related work and section III will introduce the materials and methods. Then, section IV will describe the experimental results. Section V will be used to discuss the results and, finally, we conclude in section VI summarizing the main conclusions and limitations.

II. Related Work

Psychological well-being is a complex construct that, according to different authors [1], [2], consists of three main dimensions: evaluative, hedonic and eudaimonic. Among these, the first one is related to the cognitive-judgmental aspect. The second one would be focused on the affective aspects, and covers feelings like happiness or sadness, and the last one would be centered on life purpose.

The expansion over last decades of Information Technologies and Communications in general, and the Internet in particular, has fostered the interest in the potential impact that these might have on psychological wellbeing. The evidence in this regard is mixed. Even though the initial studies identified an inverse association [3] subsequent ones questioned those results. Among these, some

Please cite this article as:
A. Cervantes, D. Quintana, Y. Sáez, P. Isasi, "Longitudinal Segmented Analysis of Internet Usage and Well-Being Among Older Adults", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 6, pp. 168-176, 2024, http://dx.doi.org/10.9781/ijimai.2022.05.002

suggested that connection might be weaker [4] or irrelevant [5]. Conversely, other studies [6], [7] report that using the Internet contributes positively to mental well-being.

The body of literature on the impact of Internet use among older adults is expanding [7]-[9]. Regarding psychosocial benefits, Forsman and Nordmyr [10] suggest that, in later adulthood, these might fall into three main categories: improved access to resources; empowered social inclusion and better interpersonal interaction.

Internet might function as a source of entertainment, which according to [8] has a direct connection to well-being among older-adults. Studies like [11] suggest that it might also double as a facilitator, fostering engagement with other activities. There is, however, a third possibility that might be especially relevant for older adults. According to some studies [12], [13], loneliness and social isolation are emerging risks in this population segment. For this reason, identifying tools to mitigate or solve these problems has the potential to have a major social impact. Having said that, the existence of positive features of digital literacy in this regard is a matter that is still under debate [14].

The most closely related study [15] explores the connection between the mentioned three dimensions of psychological well-being and Internet use in older adults relying on evidence gathered by the English Longitudinal Study of Aging (ELSA) [16].

This project has been collecting data on a bi-yearly basis from a representative sample of the English population aged 50 and older since 2002 to gain a better understanding of the ageing process. The study tracks a broad range of items that include aspects related to physical and mental health, economic position, or social participation, among others. The project is related to similar studies like the US Health and Retirement Study (HRS), the Survey of Health, Ageing and Retirement in Europe (SHARE) or the Japanese Study of Aging and Retirement (JSTAR), to mention a few.

According to these results, the connection between the main predictor and the scores on the scale used to measure the eudaimonic aspect was positive and statistically significant. However, they did not support the relevance of digital literacy on the evaluative and the hedonic components of psychological well-being.

This begs the question of whether the latter lack of connection is real or might be explained by the fact that the analysis considers the population to be homogeneous when that might not be case. Therefore, in this paper we perform a segmented analysis based on clusters identified using Self-Organizing Maps [17] instead of a global one; the objective is to confirm whether the conclusions in [15] are valid for the complete population or should be evaluated on the light of this segmentation. This research will contribute to gain a better understanding on how belonging to each of these groups impacts the connection between Internet literacy and psychological well-being. Finally, as users transition between groups over successive waves of the study, we shall also be able to perform an analysis on the evolution of users in the population.

III. Materials and Methods

A. Study Population

The data analyzed in this piece of research matches the one used in a previous study [15]. They were originated from the English Longitudinal Study of Aging. This survey tracks on a bi-yearly basis the evolution of aging and quality of life among older people in England. The sample covers community-dwelling population aged 50 and over. Among the wide range of aspects that included in the core component, we could mention health-related items; household and individual demographics; social participation; income and assets; expectations etc. Some waves supplement this information with one-off modules and questions.

The study that we present is based on data from waves 3 to 7, that is, the interviews were carried out in 2006-07, 2008-09, 2010-11, 2012-13 and 2014-15. At wave 3, 9,771 individuals took part in the study. Out of these, we lacked complete baseline data for 2,814, who were removed. Between waves 3 and 7, many individuals were lost to follow up. This further reduced the sample size to 3,547 participants by wave 7. Out of these, the information was complete over the five waves for 2,314 individuals over 50, who comprise the final dataset. We provide a table that summarizes the demographics in the Appendix.

Attrition was associated with less educated individuals that tended to show a higher degree of functional impairment. These participants, who were slightly older, also reported lower net wealth and lower digital literacy.

B. Measurements

As we mentioned, the study considers the three core dimensions of psychological well-being together with an indicator of internet literacy. The former was quantified using the indicators suggested in [18], and the second with a specific questionnaire item.

Evaluative well-being was evaluated using the Satisfaction with Life Scale (SWLS) [19]. This indicator is defined in the range 0-30, being the higher values associated with the greatest satisfaction with life. In regard to hedonic well-being, it was measured using the Enjoyment of Life Scale (EOLS). This indicator, already used in other studies [20], [21], is defined in the range 0-12. Once again, higher scores are associated with higher enjoyment of life. The instrument used to measure eudaimonic well-being consists of the items of CASP-19 not considered in EOLS (EDS). The minimum score is 0 and the maximum 45. As in EOLS, the higher the score, the greater is the eudaimonic well-being. Finally, the study proxies internet literacy through the answer to the question “I use internet/E-mail: yes/no”. The resulting dichotomous variable was encoded using (0) for negative responses and (1) for positive ones.

In addition to these measurements, the analysis also controlled for socioeconomic and health indicators. The former included age, whether the individual was a woman (1) or a man (0) and the highest academic qualification using a 3-way split to represent whether the participant had no qualification, an intermediate one, or a degree or equivalent. Given that marriage tends to be associated with more well-being [22], [23], we also considered whether the participant was legally married (1) or not (0). This was complemented with net non-pension household wealth, identified as important in previous studies [24], a dichotomous variable that indicates whether the subject volunteered (1) during the previous year or not (0), relevant according to [25], [26] and associative interests. This proxy for social connectedness, identified to have an impact on psychological well-being in studies like [26], measured the diversity of organizations that the subject reported to be part of among eight possible broad categories including, among others political parties, religious groups.

Physical activity: self-reported physical activity encoded using four consecutive levels, from 0 to 3, that represent the categories sedentary, low, moderate, and high, respectively. Finally, it was considered high in case it involved heavy manual work or vigorous activity more than once a week. The study also considers whether the participant reported having suffered limitations in instrumental activities of daily living (IADSLs) or activities of daily living (ADLs) caused by mental, physical, memory or emotional problems for a period over three months (0) or not (1). Finally, the instrument used to measure cognitive ability was the learning recall test included in the Consortium to Establish a Registry for Alzheimer Disease (CERAD) Neuropsychological battery [27]. Here, higher scores are associated with higher cognitive abilities.

Table I reports the main descriptive statistics for all these baseline characteristics of the analytical sample measured at Wave 3 [16].

The factors and covariates were measured at all follow-up interviews except for age, sex, education, wealth, and marital status, which were only measured at baseline.

TABLE I. Baseline Characteristics of the Analytical Sample Measured at Wave 3. English Longitudinal Study of Aging 2006-07. Main Descriptive Statistics

	Mean	Std. Dev.	Min.	Max.
SWLS Score*	20.49	6.159	0	30
EOLS Score*	10.13	1.624	2	2
EDS Score*	32.94	6.589	6	45
Internet/Email User	0.66	0.475	0	1
Delayed Recall	5.32	1.778	0	10
Physical Activity	2.09	0.719	0	3
Org. membership	1.79	1.410	0	8
Voluntary Work	0.39	0.487	0	1
Age	61.62	7.668	50	99
Sex	0.55	0.498	0	1
Marital Status	0.75	0.431	0	1
Education	1.18	0.807	0	2
Lack of impair.	0.85	0.360	0	1
Net wealth**	412.7	785.9	-51.9	20.818

* Scales used to measure the tree core components of psychological wellbeing: evaluative (SWLS), hedonic (EOLS) and eudaimonic (EDS).

** Net wealth in thousands of pounds.

C. Analytical Approach

The analytical strategy followed in this study combines three different instruments that will be used sequentially. The initial step will be clustering the participants according to drivers of psychological well-being using Self Organizing Maps. Then, the influence of Internet/Email use on the three main dimensions of psychological well-being by group will be assessed generalized estimating equations. Finally, the dynamic aspects of transitions among clusters will be explored using Markov models.

D. Clustering Based on Self-Organizing Maps

Self-Organizing Maps (SOM) [17], [28] is a type or artificial neural network frequently used to perform clustering analysis. This method adapts a grid of "neurons" to a specific topology. These grids are a powerful representation technique because the multidimensional topology of data can be projected in the two-dimensional space defined by the relative positions of neurons on the grid. After training, each neuron is characterized by a set of features (codebook) that summarize the features of data in its vicinity.

SOM has been used in many applications due to its simplicity and accuracy in unsupervised learning. Its learning rule can be more efficient than competing network architectures for large datasets and high dimensionality. SOM can be quite efficient to perform an unsupervised preprocessing step that is later classified by another algorithm. For instance, SOM has been used to reduce noise and dimensionality on protein classification problems [29]. The output of this phase was processed by a Particle Swarm classification. Also, it has been proved useful as part of a methodology in Human Sentiment Classification [30]. In that work SOM is used to cluster the initial patterns into separate groups, that are later classified using a Convolutional Neural Network. This approach improves the accuracy of classification over the competing techniques.

The Self Organizing map is a discrete interconnecting network (that is, a map) of neurons (also called "units"). This map is adapted to a provided training set by minimizing a loss function, the quantization error. The learning process of SOM can be summarized as follows:

First, the map is initialized. Original SOM used random initialization, but recent versions typically use Principal Component Analysis for this task. Each input neuron has N inputs, as many as features as inputs. A vector of numerical weights w_i is associated to each neuron. Thus, each neuron has a representation in the N-dimensional input space.

Upon initialization, neurons are interconnected in a specific way (for example a two-dimensional matrix), where grid distances can be measured to gauge the degree of influence among neurons.

During training, the following adaptation process is applied for each input vector x.

1. Determine the best matching neuron (BMU), as the one with the minimum Euclidean distance to the input vector (1):

$$BMU = argmin_i \, || \, x - w_i \, ||^2 \quad (1)$$

2. Adjust the weights of the BMU and also some or all the neurons in the map are "moved" closer to **x**; this is summarized using the following learning rule for all neurons *j* (2):

$$w'_j = w_j + \eta \, H \, (BMU, j) \, (x - w_j) \quad (2)$$

3. The process is repeated for all the training values over a certain number of epochs or iterations.

In the learning rule above, η is a small learning rate that is used to tune the speed of the algorithm convergence. The matrix **H** is called *function of lateral interaction*, and is composed of positive numbers that determine how intense is the modification of the weights of neurons that are neighbors of the BMU on the grid. Neurons directly connected to the BMU are "dragged" more than neurons re connected at distance 2, etc. Different versions of this function **H** exist, but they all have in common that **H** is dynamic during learning. At the start of the SOM training, influence is more global: influence is significant for a certain radius of influence from the BMU; and at the end of the process, adaptation is mostly local, that is, only the BMU and maybe its direct neighbors on the grid are adapted.

One of the advantages of SOM is that is a model-based clustering method. Once trained, the SOM network weights are retained and can be saved as a model that can be used later when new patterns become available. Thus, a SOM can be trained with the first wave of data, and later the evolution of each customer's record can be followed to check if the cluster to which it is assigned changes in later waves, thus tracking its evolution over time.

Adaptation of the SOM is only the first step of the process. The resulting model is an approximate representation of the distribution on the original data, but with a much more limited number of elements. These elements are further grouped down to generate a manageable number of clusters. This latter stage relied on Hierarchical Agglomerative clustering (HAC) [31]. This method starts with as many seeds as initial elements to be clustered. Then, recursively selects the two closest seeds in terms of the desired criterion and generates a conglomerate element with averaged values for its features. The process continues until the number of elements matches the number of desired clusters, at which point the algorithm stops.

1. Cluster-Specific Analysis Using Generalized Estimating Equations

The clusters identified in the first step of the process represent the main broad categories in which we can classify the participants according to the socioeconomic and health drivers of psychological well-being introduced in section 2.2. Once the individuals were assigned to the different groups, it was possible to perform an exploratory segmented analysis with the potential to reveal dependencies that might be difficult to identify when one considers the whole population.

The second part of the analysis relied on generalized estimating equations. These models, which are closely related to generalized linear models, provide the capability to study population-averaged effects across repeated measurements [32]. This allowed us to assess the influence of digital literacy on psychological well-being by dimension and segment of the population, controlling for potentially relevant covariates.

2. Markov Models

Markov models are commonly used for analysis of the temporal dynamics, though specific techniques used depend facts and assumptions on available data. In [33], the authors presented a generic framework for this analysis. In literature the most referenced method for parameter estimation is Expectation-Maximization (EM) [34].

A characteristic of panel data is that it can be more properly described as a mixture of models, where different groups of participants show different behavior in the temporal dimension. These models are called Latent Segment Markov Chain models. For instance, in [35] authors address how market segmentation helps providing insight on the different models that apply to each segment, without prior knowledge of the segment to which specific users belong. Some reviews of usage of these techniques in analysis can be found in [36].

Markov modelling has been used for the variation in social network structures [37] or psychological evaluation of patients [38] where specific randomization techniques are introduced to take into account interpersonal variations in the population.

In general, these models are fitted to provide both quantitative predictions on the unobserved variables (model states) and qualitative descriptions of the temporal variation on data. This type of study may provide insight even in cases where data are insufficient to provide statistically reliable predictions.

HMMs were fitted and plotted using the R package seqHMM [39].

IV. Results

This section reports the experimental results. To that end, it starts describing the experimental setup. Then, it focuses on the cluster analysis and the group-specific statistical analysis. Finally, it discusses the dynamic analysis.

A. Experimental Setup

The purpose of this work is to examine in depth the relationship between Internet usage and each of the three measures of well-being. We suspect that these relationships can't be properly assessed by analyzing the joint population of participants in the study. Thus, a more detailed analysis was performed by introducing a preliminary stage that groups participants based on the values of the covariates, the aforementioned Self-Organized Map based clustering.

Generation of the SOM maps and construction of the dynamic models was performed on a standard Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz machine with 16 GB of memory.

The overall experimental procedure was a sequence of three steps, that we cover in the following sections. As a summary, these three steps were:

Cluster analysis. The values of indicator and predictor variables were removed from the data set.

The quality of a SOM can be measured using several metrics. In this analysis we used two: topographic error and quantization error. Topographic error counts the number of times that, for any given sample vector, the second-closest neuron is not located in the immediate neighborhood of the closest neuron. If this error is low, the SOM is accurately representing the topology of data. Quantization error is calculated as the average distance from each vector to the closest neuron codebook. We tested three different grid configurations (5×5, 10×10 and 15×15). We verified that as number of neurons grew, topographical error increased, while quantization error decreased. Thus, for practical considerations we have considered that an intermediate map of 10 × 10 neurons was enough to provide a topographically accurate representation ($T_{error} = 8 \cdot 10^{-4}$) with adequate quantization error ($Q_{error} = 7.43$).

The second step was to apply the HAC algorithm to generate a small number of clusters. From the cluster u-matrix representation we chose four clusters as target. The result is an assignment of all individuals in the first wave to one of these four groups.

The software used for this was SOMbrero [40] an R package. Data loading and filtering, and generation of the SOM and clusters took 15 sec in our platform.

Group-specific statistical analysis. This analysis was performed using generalized estimating equations, so we can determine whether the conclusions are dependent on the group to which customers belong.

We fitted *c*x3 models, where *c* is the number of clusters identified by SOM. For every cluster, there were three different models, each of them targeting the scores of the relevant scales as the dependent variable (SWLS for the evaluative dimension, and EOLS and EDS for the hedonic and eudaimonic ones, respectively). All models shared the main predictor and the covariates. Following [15] and other previous studies, age and wealth were stratified. The former considered the intervals 50-59, 60-69, 70-79 and >79, and the latter quintiles.

The computation of the regression coefficients of the 2-year lagged models and their associated confidence intervals relied on model-based estimations of the covariance matrices. The statistical contrast used to assess the significance con the coefficients was the Wald test.

The software package used both to fit the models and evaluate the results was SPSS 23.

Dynamic Analysis. This analysis was performed using the map generated initially for the initial wave, but applying it to the successive waves. We separated users between in two populations: those who started with an affirmative answer to the question on Internet use (Internet users), and those who answered negatively. Then, we identified the group to which each customer belonged for different study waves. Then we constructed Markov Models that represent the transitions found between successive waves for all users. This was performed independently for both possibilities of the "Internet Use" variable in the initial wave.

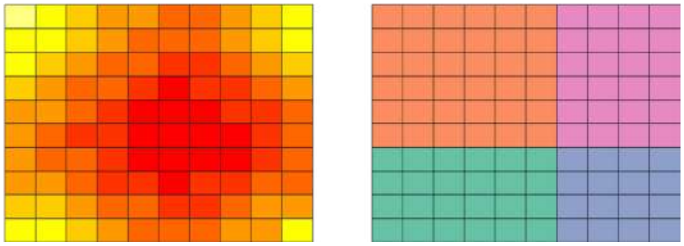
The construction of Markov models using data partitioned into groups is a mechanical task, given that we already had identified the number of possible states (4) for each customer. It is performed by calculating the conditional transition probability $P(C_{w=i+1}|C_{w=i})$ for all members of the study, and the starting $P(C_{w=0})$ for every individual.

The software used for this was self-programmed R code. Generating this model took less than 1 sec in our experimental platform.

B. Cluster Analysis

Automatic clustering was first performed on wave 3 of the available data. The resulting clusters, (hereafter called "groups") identified in this first wave are later used as reference to monitor evolution over time, by incorporating waves 4 to 7. Clustering was performed in two stages: first, a Self-Organizing Map (SOM) was used to generate a set of neurons that closely represent distribution of the covariate attributes of each participant; secondly, a hierarchical clustering algorithm was used to separate neurons in four groups. Indicator (well-being scores) and predictor variables (Internet usage) are then examined on the different groups to see if their inter-relationships are of a different nature depending on the group.

Fitting the basic SOM to the sample results in the structures illustrated in Fig. 1a. There we can see an *umatrix* representation of the SOM grid obtained for a SOM with 10 × 10 neurons. This representation depicts Euclidean distance between the neuron codebooks by using lighter colors for closer distances and darker colors for longer distances Thus, darker (red) areas can be used to separate groups of neurons that represent individuals whose features (values of the covariates in our case) are more distant.



(a) Umatrix for the 10 x 10 grid.

(b) Cluster assignments.

Fig. 1. SOM umatrix charts for wave 3 and corresponding groups after clustering.

The second step was to apply the HAC algorithm to generate a small number of clusters. Visual inspection of Fig. 1a clearly suggests the existence of 4 clusters, as four distinct areas are identified (in yellow) where neurons are grouped, separated from the rest with darker areas. Given this number as input, the HAC super-clustering process divided the map as indicated in Fig.1b, where each color represents a different cluster.

TABLE II. Averages for Each Group on Indicator and Predictor Variables, Sorted by Average Value of Internet Use

	SWLS		EOLS		EDS		Int. Use	
	Score	**Rank**	**Score**	**Rank**	**Score**	**Rank**	**Score**	**Rank**
Group 1	21.90	1	10.50	1	34.66	1	0.84	1
Group 2	20.65	3	10.26	2	33.61	2	0.77	2
Group 3	21.06	2	10.13	3	32.76	3	0.48	3
Group 4	18.53	4	9.57	4	30.41	4	0.43	4

In Table II we average and rank the values of indicator and predictor variables for each group of individuals. Group numbers have been selected in order to match the degree of Internet use from Group 1 (highest) to 4 (lowest). It is immediate to see that values for two of the three components of psychological well-being scores (EOLS and EDS) are sorted in the same descending order. Therefore, regarding these two measures of well-being, Group 1 has the highest level of psychological well-being and also the highest level of Internet use, and Group 4 has the lowest level for both measures.

However, this generic correlation is not true for all groups regarding SWLS: in this case, Group 3 ranks second in the SWLS score, over Group 2 which is third, though differences are not great.

Average Internet use in Groups 1 and 2 is very similar, and there is also a small difference between Groups 3 and 4. However, well-being scores do not show such division, and present a more gradual distribution.

In order to evaluate the properties of each of the groups, we have averaged the values of all covariates for each cluster in Table III. Here we see that some of the covariates follow the same ordering and might be equally correlated to psychological well-being: Cognitive ability (Delay Recall), qualification (Education) and net non-pension income (Net Wealth). Others, such as degree of membership to organizations and Volunteer work seem to account for much of the distinction between clusters 1 and 2, and also between clusters 3 and 4.

C. *Group-Specific Statistical Analysis*

The statistical analysis results are summarized in Table IV. There, we report the main coefficients of the 12 GEE models, one per combination of cluster and scale, together with the associated p-values obtained using the Wald test. These represent the differential impact of digital literacy on the results of the three scales used to proxy the three key dimensions, controlling for all the covariates, as discussed in 2.3.2.

If we focus our attention on those that are statistically significant at the 5% conventional level, the use of Internet/Email had a generally positive impact. However, the beta coefficients of the 2-year lagged models reveal the existence of differences among clusters. This is especially noticeable in cluster 3, as the model suggests an inverse relationship between Internet literacy and the hedonic dimension.

TABLE IV . Summary of GEE Analysis of Connection Between Internet Use and Psychological Well-Being Test Scores by Cluster, English Longitudinal Study of Aging 2006-14

	SWLS Score[a]		EOLS Score[a]		EDS Score[a]	
	Coeff.[b]	**P[c]**	**Coeff.[b]**	**P[c]**	**Coeff.[b]**	**P[c]**
Group 1	0.24	.45	0.12	.17	0.44	.22
Group 2	0.47	.06	0.04	.59	0.60	.03
Group 3	-0.29	.57	-0.24	.002	-0.25	.43
Group 4	0.50	.04	0.07	.006	0.87	.001

[a] Scales used to measure the tree core components of psychological well-being: evaluative (SWLS), hedonic (EOLS) and eudaimonic (EDS).
[b] Beta regression coeffcients estimated through 2-year lagged generalized estimating equations.
[c] P values from Wald test.

D. *Dynamic Analysis*

In this section we are concerned by how individuals migrate from one group to another during the successive waves of the study.

In Table V we show the variation on the population in the formerly calculated groups. These figures were calculated on the individuals

TABLE III. Average Values of Covariates by Group. Group 1 Has the Highest Average Internet Use, and Group 4 the Lowest

	N[a]	Age[b]	Sex[c]	MrS.[d]	Edu.[e]	NW[f]	Imp.[g]	Phys.[h]	Rec.[i]	Org.[k]	Vol.[l]
Group 1	590	59.81	0.55	0.84	1.68	624.1	0.94	2.38	6.12	2.93	0.94
Group 2	798	58.51	0.43	0.83	1.41	419.0	0.96	2.39	5.66	1.46	0.00
Group 3	336	68.21	0.61	0.68	0.99	326.5	0.71	1.83	4.32	2.09	0.93
Group 4	590	63.86	0.68	0.59	0.49	241.6	0.67	1.56	4.65	0.94	0.03

[a] Participants; [b] Age; [c] Sex; [d] Marital status; [e] Education; [f] Net Wealth (Thousands of pounds); [g] lack of impairments; [h] Physical activity; [i] Delayed recall; [k] Organization membership; [l] Voluntary work

that were present in the five waves of the study. It is obvious that ageing must have an overall impact that is easily shown in the total number of individuals that compose each of the groups. The number of participants in Group 3 increases by a factor of 1.76 between wave 3 and wave 7, and Group 4 increases by a factor of 1.34. On the other hand, participants in groups of younger average age decrease with time: Group 1 decreases by factor of 0.65, and Group 2 by 0.68.

TABLE V. Evolution of the Number of Individuals Per Cluster. Groups Are Calculated for Wave 3. For Other Waves, Each Individual Is Assigned to the Group to Which the Closest Codebook in the SOM Grid Belongs

	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7
Group 1	590	522	498	455	385
Group 2	798	749	649	587	543
Group 3	336	401	471	528	593
Group 4	590	642	696	744	793

As we can see in Table VI, with 2314 participants, we have 9256 possible transitions. The diagonal totals 6374, that is in 68.86% of the cases individuals do not change the group to which they belonged to at the start of the study.

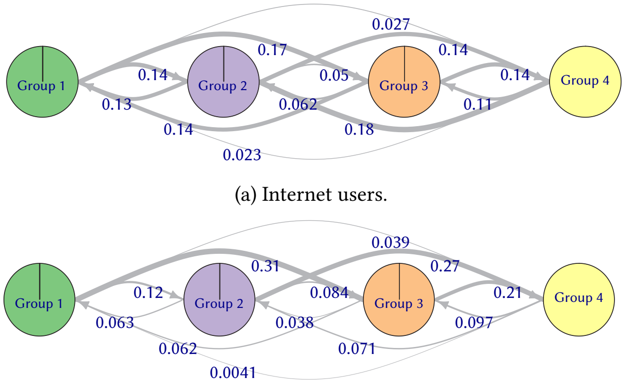
TABLE VI. Transition Table in Number of Cases. In Rows, the Starting Group (in Any Wave From 3 to 6); in Columns the Destination Group in the Following Wave (4 to 7)

	Group 1	Group 2	Group 3	Group 4
Group 1	1328	288	389	60
Group 2	321	1828	159	475
Group 3	177	89	1174	296
Group 4	34	323	271	2044

This information may be used to construct a transition table that details the probability of a user to either move to, or stay in any group, depending on the original group. This is equivalent to constructing two independent first-order Markov chain models of group labels.

We have constructed the global Markov chain for group transitions using the classical EM method [34] on the full chains of four elements. We must point out that these figures are only approximate, as we are averaging results for the four transitions between different waves that can't be modelled as time-homogeneous.

The resulting transition probabilities are shown graphically in Fig. 2. In this figure, arrow width is a measure of the transition probability value.



(a) Internet users.

(b) Non-internet users.

Fig. 2: Cluster transition diagrams. Group 1 has the highest scores in wellbeing, while Group 4 has the lowest. Transition probabilities printed on edges.

For clarity, in Table VII and VIII we show the transition probabilities in these figures.

TABLE VII. Transition Probabilities for Internet Users. In Rows, the Starting Group (in Any Wave From 3 to 6); in Columns the Destination Group in the Following Wave (4 to 7)

	Group 1	Group 2	Group 3	Group 4
Group 1	0.662	0.143	0.168	0.027
Group 2	0.130	0.678	0.050	0.142
Group 3	0.135	0.062	0.664	0.139
Group 4	0.023	0.182	0.106	0.688

TABLE VIII. Transition Probabilities for Non-Internet Users. In Rows, the Starting Group (in Any Wave From 3 to 6); in Columns the Destination Group in the Following Wave (4 to 7)

	Group 1	Group 2	Group 3	Group 4
Group 1	0.533	0.121	0.307	0.039
Group 2	0.063	0.580	0.084	0.237
Group 3	0.062	0.038	0.692	0.208
Group 4	0.004	0.071	0.097	0.828

V. Discussion

The SOM analysis resulted in the identification of four groups of participants whose main characteristics were summarized in Table III. The first one showed high values for all the well-being scores and the highest internet use. This group had the second lowest average age. It showed a high level of physical activity that corresponded to the lack of impairments. This group also showed high level of participation in organizations. The second group had a high use of internet and high values for all the well-being scores except for the SWLS score. Level of activity and lack of impairments levels were the same as in group 1. However, there was a clear difference in net wealth, level of organization membership and voluntary work. There was also a preponderance of males compared to group 1. Group 3 had the highest average age, and covariates such as marriage status, education, physical activity, and net wealth were lower than the previous groups. It also showed a much lower internet use. Conversely, it was characterized by higher organizational membership and voluntary work. Finally, the fourth group was characterized by the least internet use and low well-being scores, even though average age was lower than members of group 3. This group was also characterized by the lowest education and physical activity levels, the highest preponderance of females, and lowest score in marital status. This group shows a very low degree of membership in organizations and a very low score in voluntary work.

The results of the GEE models supported the basic hypothesis: studying the population as a whole causes a loss of relevant information vs. a more fine-grained segmented analysis.

Internet literacy does not seem to have any significant connection with psychological well-being for the participants in group 1. The rest of the groups present different associations with the different scales of well-being:

Internet use seems to have an association with the SWLS score for groups 2 and 4, though this is only significant at 5% for group 4.

For the eudeimonic (EDS) score, this connection is significant at 1% for both groups 2 and 4.

Results for the hedonic (EOLS) score were particularly interesting. We found a positive association from Internet use to EOLS for group 4. However, the most interesting result in this regard is the existence of a clear negative relationship with enjoyment of life for group 3.

There are some differences with results obtained by Quintana et al. when analyzing the whole sample [15]: first, in that work neither SWLS nor the EOLS score could be associated with Internet use at conventional levels, whereas we have found groups for which that association can be significant; secondly, the significant association of Internet use and the score in the EDS scale reported in the work on the aggregated data is now shown to be related (and significant at 1%) to participants in groups 2 and 4 (approximately 60% of the sample); finally, the negative association in group 3 was not detected when considering the whole sample.

The dynamics captured by the Markov models show markedly differentiated transition probabilities among groups depending on digital literacy. In Fig. 2, arrows from left to right mean (in general) decrease in the wellbeing scores. Those transitions are much more balanced for Internet users: internet users are more likely to move in both directions, where non internet users are more likely to decrease their levels of well-being over time. In fact, non-Internet users have an 83% probability of staying in the lowest-scoring group (4) once they fall in it, and their probabilities of reaching that state from groups 2 and 3 are much higher.

The findings derived from both the static and the dynamic analysis open the door for more targeted research that could provide more insights on the connection between Internet use and well-being and implementation of better targeted intervention programs.

Current literature is yet to reach a definitive conclusion on the relation between aspects of psychological well-being and Internet literacy among the older population [14]. There are several important problems to common approaches, both from the methodological and from the data collection point of view, especially in terms of the impacts of different types of Internet use [8]. Our work confirms that segmentation of data may provide significant insights that help comprehend seemingly contradictory results. For instance, some studies report positive relationships between these variables: use of Internet and self-reported life satisfaction in [7], or with the hedonic dimension of well-being [6], as it reduces the probability of depression very significantly. On the other hand, negative or mixed relationships among certain uses of the Internet and perception of wellbeing have also been documented. In [41] authors use of Internet for communication with unknown people is a symptom of feeling of loneliness, while communication with family members reduces those feelings. Also [42] points out that frequent use of Internet may have a positive association with depression.

We must point out some challenges of this study that have to do with the representativity of the sample. Our data is based on ELSA and thus has been recorded on various geographical locations in England. Evidence from previous literature suggest that most research conclusions on ageing studies can only be generalized to countries on with similar levels of development. What is more, some studies on aging point out differences among results from European elderly population and American counterparts [43]. The sample included participants aged 50 years or older not living in assisted living or nursing homes, and those speciffic living conditions might not be directly extrapolated to the general population.

Finally, we will point out a limitation regarding data availability. Most of the waves considered in this study provide very limited information on the type of use of Internet: only the dichotomous response used as independent variable is available. More details on intensity and specific uses would leave room for more detailed analysis, along the lines the one described by Hofer et al. [44] on online information seeking. We hope that we will be able to go deeper into the analysis and get a clearer picture once data on new waves gets released over the next years.

VI. Conclusions

This study provides new insights on the connection between Internet use and three core dimensions of psychological well-being at advanced age.

The results support our two initial hypothesis: the existence of a segmented population in terms of the main drivers of well-being, and the importance of performing a fine-grained segmented analysis.

The existence of four clusters and the differential impact of digital literacy depending on the group opens the door to further research and the development of specific interventions. The latter is especially relevant in the light of the dynamic analysis, as there seems to be a clear association between this factor and the potential for transition to segments of the population characterized by higher levels of psychological well-being.

From an instrumental point of view, the results support a high potential for Self-Organizing Maps and Markov models in this domain.

Appendix

Baseline Characteristics of the Analytical Sample by Psychological Well-Being Indicator. English Longitudinal Study of Aging 2006–2007

	n (%)	SWLS Score *		EOLS Score *		EDS Score *	
		Mean	Std.	Mean	Std.	Mean	Std.
Internet/Email User							
No	792 (34%)	19.77	6.368	9.93	1.781	31.20	6.983
Yes	1522 (66%)	20.86	6.015	10.24	1.526	33.59	6.280
Age							
50–59	1059 (46%)	19.58	6.559	10.05	1.708	32.74	6.721
60–69	839 (36%)	20.82	5.983	10.20	1.539	33.28	6.719
70–79	385 (17%)	21.01	5.355	10.16	1.577	32.62	6.076
>79	31 (1%)	22.52	4.434	10.03	1.791	34.10	6.156
Sex							
Male	1041 (45%)	20.64	5.934	10.04	1.648	32.85	6.370
Female	1273 (55%)	20.36	6.336	10.12	1.600	33.01	6.765
Education							
None	579 (25%)	20.11	6.220	9.94	1.699	31.65	6.984
Intermediate	729 (32%)	19.98	6.391	10.02	1.640	32.78	6.639
Degree	1006 (43%)	21.07	5.904	10.22	1.625	33.79	6.183
Lack of Impairments							
Yes	354 (15%)	16.64	7.169	9.15	1.806	28.27	7.152
No	1960 (85%)	21.05	5.786	10.31	1.523	33.78	6.115
Marital Status							
Single	572 (25%)	18.05	6.896	9.86	1.659	32.49	6.884
Married	1742 (75%)	21.29	5.676	10.22	1.602	33.08	6.485
Physical Activity							
Sedentary	42 (2%)	16.57	7.979	8.86	1.555	27.51	8.055
Low	373 (16%)	19.34	6.709	9.65	1.750	30.50	7.065
Moderate	1225 (53%)	20.48	5.975	10.12	1.614	32.99	6.331
High	674 (29%)	21.37	5.849	10.49	2.125	34.52	6.099
Voluntary Work							
No	423 (61%)	19.80	6.438	9.99	1.689	32.34	6.875
Yes	891 (39%)	21.59	5.511	10.35	1.489	33.89	5.988
Wealth Quintile †							
Q1	290 (12%)	17.27	7.517	9.42	1.863	29.48	7.655
Q2	383 (16%)	19.34	6.539	9.80	1.728	31.24	7.075
Q3	479 (21%)	20.38	5.919	10.19	2.307	32.51	6.360
Q4	548 (24%)	21.27	5.393	10.34	1.497	33.73	5.896
Q5	614 (27%)	22.10	5.260	10.43	1.489	35.25	5.379
Org. membership **Delayed Recall**	2314 (100%)	20.49	6.159	10.13	1.624	32.94	6.589

[*] Scales used to measure the tree core components of psychological well-being: evaluative (SWLS), hedonic(EOLS) and eudaimonic (EDS).

[†] Quintile distribution based on the initial unfiltered sample, not the analytical one. Higher quartiles represent more wealth.

Acknowledgment

The data were made available through the UK Data Archive. ELSA was developed by a team of researchers based at the National Centre for Social Research, University College London and the Institute for Fiscal Studies. The data were collected by the National Centre for Social Research. The funding is provided by the National Institute of Aging in the United States and a consortium of UK government departments coordinated by the Office for National Statistics. The developers and funders of ELSA and the Archive do not bear any responsibility for the analyses or interpretations presented here.

This work has been supported by the Madrid Government (Comunidad de Madrid-Spain) under the Multiannual Agreement with UC3M in the line of Excellence of University Professors (EPUC3MXX), and in the context of the V PRICIT (Regional Programme of Research and Technological Innovation).

References

[1] D. Kahneman and A. Deaton, "High income improves evaluation of life but not emotional well-being," In *Proceedings of the National Academy of Sciences of the United States of America,* vol. 107, pp. 16489–93, 2010.

[2] P. Dolan, R. Layard and R. Metcalfe, "*Measuring Subjective Wellbeing for Public Policy: Recommendations on Measures*," CEP Special Papers Centre for Economic Performance, LSE, 2011.

[3] R. Kraut, M. Patterson, V. Lundmark, S. Kiesler, T. Mukopadhyay and W. Scherlis, "Internet paradox: A social technology that reduces social involvement and psychological well-being?," *The American psychologist,* vol. *53*, pp. 1017–31, 1998.

[4] R. Kraut, S. Kiesler, B. Boneva, J. Cummings, V. Helgeson and A. Crawford, "Internet paradox revisited," *Journal of Social Issues,* vol. *58*, pp. 49–74, 2002.

[5] L. A. Jackson, A. von Eye, G. Barbatsis, F. Biocca, H. E., Fitzgerald and Y. Zhao, "The impact of internet use on the other side of the digital divide," *Communications of the ACM,* vol. 47, pp. 43–47, 2004.

[6] S. R. Cotten, G. Ford, S. Ford and T. M. Hale, "Internet use and depression among older adults," *Computers in Human Behavior,* vol. 28, pp. 496–499, 2012.

[7] O. Lelkes, "Happier and less isolated: internet use in old age," *Journal of Poverty and Social Justice,* vol. 21, pp. 33–46, 2013.

[8] R. Lifshitz, G. Nimrod and Y. G. Bachner, "Internet use and well-being in later life: a functional approach," *Aging & Mental Health,* vol. *22*, pp. 85–91, 2018.

[9] B. B. Neves, R. Franz, R. Judges, C. Beermann and R. Baecker, "Can digital technology enhance social connectedness among older adults? A feasibility study," *Journal of Applied Gerontology,* vol. 38, no. 1, pp. 49–72, 2017.

[10] A. K. Forsman and J. Nordmyr, "Psychosocial links between internet use and mental health in later life: A systematic review of quantitative and qualitative evidence," *Journal of Applied Gerontology*, vol. 36, pp. 1471–1518, 2017.

[11] M. Näsi, P. Räsänen and O. Sarpila, "ICT activity in later life: Internet use and leisure activities amongst senior citizens in Finland," *European Journal of Ageing,* vol. 9, pp. 169–176, 2012.

[12] N, Savikko, P. Routasalo, R. Tilvis, T. Strandberg and K. Pitälä, "Predictors and subjective causes of loneliness in an aged population," *Archives of Gerontology and Geriatrics,* vol. 41, pp. 223–233, 2005.

[13] P. A. Dykstra, T.G. van Tilburg and J. de Jong Gierveld, "Changes in older adult loneliness: Results from a seven-year longitudinal study," *Research on Aging,* vol. *27*, pp. 725–747, 2005.

[14] R. Beneito-Montagut, Y de, N. Cassián and A. Begueria, "What do we know about the relationship between internet mediated interaction and social isolation and loneliness in later life?," *Quality in Ageing and Older Adults*, vol. 19, no. 1, pp. 14–30, 2018.

[15] D. Quintana, A. Cervantes, Y. Sáez and P. Isasi, "Internet use and psychological well-being at advanced age: Evidence from the English longitudinal study of aging," *International Journal of Environmental Research and Public Health,* vol. 15, no. 3, art. 480, 2018.

[16] J. Banks, G. D. Batty, K. Coughlin, P. Dangerfield, M. Marmot, J. Nazroo, Z. Oldfield, N. Steel, N., Steptoe, M. Wood and P. Zaninotto, "English Longitudinal Study of Ageing: Waves 0-9, 1998-2019". [data collection]. 33rd Edition. UK Data Service. SN: 5050, 2019, http://doi.org/10.5255/UKDA-SN-5050-20

[17] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics,* vol. 43, pp. 59–69, 1982.

[18] A. Steptoe, P. Panayotes and C. de Oliveira, "The psychological wellbeing, health and functioning of older people in England," [in] the dynamics of ageing: evidence from the English longitudinal study of ageing 2002-10 (Wave 5). In *The dynamics of ageing: evidence from the English longitudinal study of ageing 2002-10 (Wave 5)* (pp. 98–182). Institute for Fiscal Studies, 2012.

[19] E. Diener, R. A. Emmons, R. J. Larsen and S. Griffin, "The satisfaction with life scale," *Journal of Personality Assessment,* vol. 49, pp. 71–75, 1985.

[20] A. Steptoe and J. Wardle, "Positive affect measured using ecological momentary assessment and survival in older men and women," *Proceedings of the National Academy of Sciences*, vol. 108, pp. 18244–18248, 2011.

[21] A. Steptoe, P. Demakakos, C. de Oliveira and J. Wardle, "Distinctive biological correlates of positive psychological well-being in older men and women," *Psychosomatic medicine,* vol. 74, pp. 501–8, 2012.

[22] M. L. Diener and M. B. Diener McGavran, "What makes people happy?: A developmental approach to the literature on family relationships and wellbeing," In *The science of subjective well-being* (pp. 347–375). Guilford Press, 2008.

[23] D. Carr, V. A. Freedman, J. C. Cornman and N. Schwarz, "Happy marriage, happy life? Marital quality and subjective well-being in later life," *Journal of marriage and the family,* vol. *76*, pp. 930–948, 2014.

[24] V. Lorant, D. Delière, W. Eaton, A. Robert, P. Philippot, P. and M. Ansseau, "Socioeconomic inequalities in depression: A meta-analysis," *American Journal of Epidemiology,* vol. 157, pp. 98–112, 2003.

[25] F. Borgonovi, "Doing well by doing good. the relationship between formal volunteering and self-reported health and happiness," *Social Science & Medicine*, vol. 66, pp. 2321–2334, 2008.

[26] A. M. Creaven, A. Healy and S. Howard, "Social connectedness and depression: Is there added value in volunteering?," *Journal of Social and Personal Relationships,* vol. 35, no. 10, pp. 1400–1417, 2017.

[27] J. C. Moms, A. Heyman, R. C. Mohs, J. P. Hughes, G. van Belle, G. Fillenbaum, E. D. Mellits and C. Clark, "The consortium to establish a registry for alzheimer's disease (CERAD). Part I. clinical and neuropsychological assessment of Alzheimer's disease," *Neurology,* vol. *39*, pp. 1159–1159, 1989.

[28] T. Kohonen, M. Schroeder and T. S Huang, T., *Self-Organizing Maps*. Springer-Verlag Berlin Heidelberg, 2001.

[29] M. S. Kamal, M. G. Sarowar, N. Dey, A. S. Ashour, S. H. Ripon, B. K. Panigrahi and J. M. R. S. Tavares, "Self-organizing mapping based swarm intelligence for secondary and tertiary proteins classification," *International Journal of Machine Learning and Cybernetics,* vol. 10, pp 229–252, 2019.

[30] M. Ali, M.G. Sarowar, M. Rahman, M. L. Rahman, J. Chaki, N. Dey and J. M. R. S. Tavares, "Adam Deep Learning with SOM for Human Sentiment Classification," *International Journal of Ambient Computing and Intelligence,* vol. 10, pp. 92–116, 2019.

[31] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika,* vol. *32*, pp. 241–254, 1967.

[32] K. Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika,* vol. *73*, pp. 13–22, 1986.

[33] J. D. Kalbfleisch and J. F. Lawless, "The analysis of panel data under a Markov assumption," *Journal of the American Statistical Association,* vol. 80, pp. 863–871, 1985.

[34] 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,* vol. 39, pp. 1–38, 1977.

[35] J. G. Dias and J. K. Vermunt, "Latent class modeling of website users' search patterns: Implications for online market segmentation," *Journal of Retailing and Consumer Services*, vol. 14, pp. 359–368, 2007.

[36] J. K. Vermunt, "Longitudinal research using mixture models," In *Longitudinal Research with Latent Variables* (pp. 119–152). Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[37] T. Snijders *Network Analysis, Longitudinal Methods of* Analysis. Springer Verlag New York, 2013.

[38] S. de Haan-Rietdijk, P. Kuppens, C. S. Bergeman, L. B. Sheeber, N. B. Allen and E. L Hamaker, "On the use of mixed Markov models for intensive longitudinal data," *Multivariate Behavioral Research,* vol. 52, pp. 747–767, 2017.

[39] J. Helske and S. Helske, *Mixture hidden Markov models for sequence data: the seqHMM package in R.* R package version 1.0.8-1.

[40] N. Villa-Vialaneix, J. Mariette, M. Olteanu, F., Rossi, L. Bendhaiba and J. Boelaert (2018). *SOMbrero: SOM Bound to Realize Euclidean and Relational Outputs.* R package version 1.2-3, 2018.

[41] S. Sum, R. M. Mathews, I. Hughes and A. Campbell, "Internet use and loneliness in older adults," *Cyberpsychology & behavior: the impact of the Internet, multimedia and virtual reality on behavior and society,* vol. 11, pp. 208–11, 2008.

[42] C. Morrison and H. Gore, "The relationship between excessive internet use and depression: A questionnaire-based study of 1,319 young people and adults," *Psychopathology,* vol. 43, no. 2, pp. 121–126, 2010.

[43] B. van Landeghem, "A test for the convexity of human well-being over the life cycle: Longitudinal evidence from a 20-year panel," Journal of Economic Behavior & Organization, vol. 81, pp. 571–582, 2012.

[44] M. Hofer, E. Hargittai, M. Büchi and A. Seifert, "Older adults' online information seeking and subjective well-being: The moderating role of internet skills," *International Journal of Communication* vol. 13, pp. 4426–4443, 2019.

Alejandro Cervantes

Alejandro Cervantes graduated as Telecommunications Engineer at Universidad Politecnica of Madrid (Spain), in 1993. He received his PhD in Computer Science at Carlos III of Madrid in 2007. He is currently an assistant professor at the Computer Science Department at this same University. His current interests focus on algorithms for classification of non-stationary data, large multi-objective optimization problems, swarm intelligence algorithms for data mining and AI systems for steganalysis.

David Quintana

Holds Bachelor degrees in Psychology and Computer Science. He has an M.S. in Intelligent Systems from Universidad Carlos III de Madrid and a Ph.D. in Finance from Universidad Pontificia Comillas (ICADE). He is currently an Associate Professor with the Department of Computer Science at Universidad Carlos III de Madrid. There, he is part the bio-inspired algorithms group EVANNAI. His current research interests are mainly focused on applications of computational intelligence in finance and education.

Yago Sáez

Received the degree in computer engineering in 1999. He got his Ph.D. in Computer Science from the Universidad Politécnica de Madrid, Spain, in 2005. Since 2007 till 2015 he was vice-head of the Computer Science Department from the Carlos III University of Madrid, where he got a tenure, and is nowadays associate professor. He belongs to the Evolutionary Computation, Neural Networks and Artificial Intelligence research group (EVANNAI).

Pedro Isasi

Graduate and Doctor in Computer science by the Polytechnic University of Madrid since 1994. Currently he is University professor and head of the Evolutionary Computation and Neural Networks Laboratory in the Carlos III of Madrid University. Dr. Isasi has been Chair of the Computational Finance and Economics Technical Committee (CFETC) of the IEEE Computational Intelligence Society (CIS), Head of the Computer Science Department and Vice-chancellor in the Carlos III University among others. His research is centered in the field of the artificial intelligence, focusing on problems of Classification, Optimization and Machine Learning, fundamentally in Evolutionary Systems, Metaheuristics and artificial neural networks.

Modulating the Gameplay Challenge Through Simple Visual Computing Elements: A Cube Puzzle Case Study

Jose Ribelles, Angeles Lopez, V. Javier Traver*

Institute of New Imaging Technologies, Universitat Jaume I, Castellón (Spain)

* Corresponding author. vtraver@uji.es

Received 24 April 2021 | Accepted 3 March 2022 | Early Access 2 May 2022

Abstract

Positive player's experiences greatly rely on a balanced gameplay where the game difficulty is related to player's skill. Towards this goal, the gameplay can be modulated to make it easier or harder. In this work, a modulating mechanism based on visual computing is explored. The main hypothesis is that simple visual modifications of some elements in the game can have a significant impact on the game experience. This concept, which is essentially unexplored in the literature, has been experimentally tested with a web-based cube puzzle game where participants played either the original game or the visually modified game. The analysis is based on players' behavior, performance, and replies to a questionnaire upon game completion. The results provide evidence on the effectiveness of visual computing on gameplay modulation. We believe the findings are relevant to game researchers and developers because they highlight how a core gameplay can be easily modified with relatively simple ingredients, at least for some game genres. Interestingly, the insights gained from this study also open the door to automate the game adaptation based on observed player's interaction.

Keywords

Computer Game, Gameplay Modulation, Player Experience, Visual Computing.

DOI: 10.9781/ijimai.2022.05.001

I. Introduction

A good understanding of the factors affecting the player experience can be very important for producing better games. Accordingly, much research effort has been devoted to both gaining theoretical insights on, and quantifying, this experience. Game heuristic and guidelines can be proposed to elicit desirable human (player) emotions and responses. Engagement is one of the concepts that has been studied, with the Game Engagement Questionnaire (GEQ) [1] being proposed to measure it. Validation studies [2] of game experience scales have been performed [3]. For game challenge, a recent scale has been developed and validated [4], [5]. Flow [6] is also widely studied in the context of video games, for better understanding it [7], [8], measuring it [9]–[13], relating it experimentally to learning and other conditions [14]–[17], or even providing some design intuitions or guidelines to produce it [8], [18].

Many of the aspects of the player experience are interrelated, so that flow, engagement [1], enjoyment [19]–[22], and others such as immersion [23], may overlap and share common attributes. For instance, flow is one of the factors within the GEQ [1]. Interestingly, most conceptualizations of the player experience share the view that the *optimal complexity level* is a key ingredient of an enjoyable game, and it is particularly essential for flow. This optimal experience relies on a balance between the challenges of the game and the player's skill; in essence, if the game is too simple or too challenging for the player, it will lead to either boring or frustrating experiences, respectively [24].

Our work has to do mainly with the game challenge; in particular, we delve into this issue: whether the player experience can be modulated with relatively simple manipulations of images present in the game. By *image manipulation* we mean modifying some other aspects of the image [25] (e.g. color, edges, texture) or distorting it somehow (e.g. blurring, geometric deformations, frequency filtering), but respecting some aspects of its contents, so that it can be recognized and potentially distinguished from others, yet with different cognitive abilities or effort with respect to the original image. This possibility can be partially grounded on the information quality, which is understood as one of the mechanisms underlying flow [8]. We generally refer to these alterations as *visual computing* [26]. Arguably, visual computing is particularly suitable for some game types or genres such as puzzle or card games, but it could also be applied to other games where the modification of visuals of some of its elements, may have an impact on the gameplay. To address this issue, we focus on a controlled case study, a 2D video game version of the Cube blocks puzzle (Fig. 1), and explore whether, and how, changes to the images used for the sides of the cubes affect the actual gameplay and the player perception of their experience.

The contributions of this work are as follows. The possibility of modulating the gameplay with visual computing elements is

Please cite this article as:
J. Ribelles, A. Lopez, V. J. Traver, "Modulating the Gameplay Challenge Through Simple Visual Computing Elements: A Cube Puzzle Case Study", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 6, pp. 177-193, 2024, http://dx.doi.org/10.9781/ijimai.2022.05.001

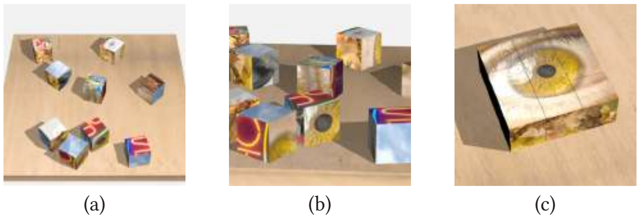


Fig. 1. A rendering of Cube blocks, a classic puzzle game: (a,b) two views of the unsolved puzzle and (c) once the puzzle is solved for a target image of a human eye. The prototype game used in this paper consists of solving a series of these puzzles, in a 2D (top-view) interface.

hypothesized. A web-based cube blocks puzzle has been developed as a proof-of-concept prototype. A user study has been conducted to find out possible behavioral, performance and opinion differences of players in one control and one experimental group. A detailed analysis of the results has been performed. Results provide interesting findings that can inform both game level designs and, eventually, automatic and dynamic game difficulty adjustment.

In the following, section II reviews related work on image manipulation and gameplay modification of some sort. Section III introduces the core game and its version with visual computing. Section IV details the design and implementation of both the experiment and the interface. Results are carefully analyzed and discussed in section V. Some discussions on the potential and limitations of the work are provided in section VI, and concise conclusions are finally presented in section VII.

II. Previous Work

In general terms, it is well accepted the role that visual contents in all kinds of media play on our emotions, and that they can even affect our beliefs. This fact has ramifications on how deliberate modifications of these contents may end up affecting our attitudes and behaviors [27]. The players' visual attention patterns can inform the level design for game improvement [28]. By masking some kinds of visual information affecting human prior knowledge, the human performance in a platform game has been found to decay notably with respect to the original unmodified game [29]. The relation between playing some games and the perceptual and cognitive abilities that are developed [30]–[34] may eventually bring insights into game design by reversing the problem, namely, how can we affect performance by modulating the gameplay. The interesting relationship between perception, computer graphics and video games are certainly not new, but with a large margin for further exploration [35].

Next, we briefly review ideas previously considered for visual or challenge modifications, which for convenience are also summarized in Table I. As a first simple choice, players may select game levels when available, although this choice not only relies on the players' ability to do the selection adequately, but also on a proper difficulty design by game developers, an area where more research and designer-assistance tools are required [36]. Similarly, modding can be seen as a form of game modification, but it is motivated more by the need of self-expression of gamers [37], than aimed at a carefully planned gameplay modification. Schell [38] mentions a few examples of subtle visual cues that can indirectly guide the player's actions with the goal of providing the player with a sense of perceived freedom, without actually enjoying full freedom. Similarly, in Mirror's Edge [39], some game elements are highlighted in red as explicit cues to navigation.

In games, the term *juiciness* refers to an abundance of audiovisual effects [40], as a form of additional feedback (more than strictly required from a usability point of view), often including second-order motion, which seeks to provide the player with plenty of power and rewards [38]. Since the term juiciness is somehow vague, a recent survey tried to elicit developers' understanding, in an attempt to provide a useful framework for the game design and research communities [41]. Regarding the positive or negative effects of "juicy" games, findings on juicy-vs-dry game versions are somehow disparate and thus results remain essentially inconclusive [42]–[47]. For instance, it has been observed that while the perceived competence is positively affected by juiciness, the actual performance is not really changed [47]. However, a recent large-scale empirical study with four levels of juiciness in a role-playing game [48] reveals that the degree of juiciness has an impact on the valence of the effects on performance, experience and motivation, with moderate amounts of juiciness found to be optimal. Although interesting and akin to our work, this form of game modification has more to do with user interaction and includes more effects (motion, audio) than the ones considered in our work.

TABLE I. Summary of (Visual) Game Modifications and Their Main Purpose

Approach	Main purpose
Level selection	Match game challenge to player's skill
Modding	Allow gamers' self-expression
Subtle visual cues	Guide players' actions
Juiciness	Endow the player with rewards and sense of power
Flipped levels	Expand the game variety
Embellishments	Increase the engagement
Non-photorealistic rendering	Provide artistic styles and moods
Color adaptations	Augment accessibility
Visual computing (proposed here)	Support gameplay (challenge) modulation

In some games, previously played levels are flipped horizontally and, optionally, other components such as the art style also changed, which is a simple means of having more levels to play; apparently these alterations can make the game harder to play [49], [50]. Embellishments in game skins are found to increase engagement but decrease performance [51]. Color in some game elements may also have some effects. Performance, immersion and flow have been found to be inferior with red avatars than with blue ones [52]. Some games include color-blindness adaptation [53], [54] some of which may result in (unintended) easier playing for people with normal vision. Non-photorealistic techniques were proposed as a means of providing the players with different styles and moods [55], but their impact in terms of gaming experience was not studied. Some past video games included effects such as toon-rendering, comic-like appearances, sumi-e painting or pointillist style [56]. Finally, procedural content generation techniques [57] can be rather sophisticated, a current research hot topic and orthogonal to our study.

Although all these are means to modify some visual aspect of games, the mechanisms and purposes are not always aimed at modulating the gameplay, as summarized in Table I. Therefore, as far as we know, using visual computing to modulate the challenge in the gameplay has scarcely been considered in the literature, and detailed studies in the sense intended in this work are limited.

To summarize, this work addresses the use of visual modification for modulating the difficulty or challenge of the gameplay by means of a case study. Notice that it is not aimed at making the players achieve the state of flow, which could be seen as an ultimate desirable goal, but it is out of the scope of this present work. Since the flow depends both on the game and on each particular player, the flow goal partially relates to the interesting concept of dynamic difficulty adjustment [36], [58]–[60], and the findings of our study can be expected to complement and inform subsequent research in this area. If we were able to characterize a player in terms of their skills on the one hand, and categorize different image modifications (including no modification at all) in terms of the degree of difficulty they induce on the other hand, then a properly balanced skill-difficulty match could be chosen for an improved game experience (Fig. 2).

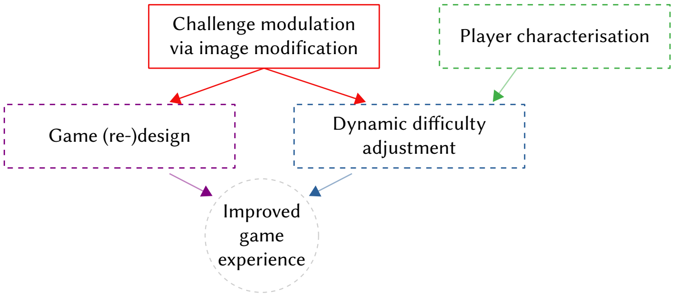

Challenge modulation via image modification
Player characterisation
Game (re-)design
Dynamic difficulty adjustment
Improved game experience

Fig. 2. This work focuses on exploring game challenge modulation through image modifications. Eventually, this might help (re-)design games for an improved game experience. Such an improvement might also be obtained together with player characterization through dynamic difficulty adjustment. Dashed lines are used to highlight that these modules are not part of this work.

III. Methodology

To evaluate the possibility of modulating the gameplay with visual computing elements, we developed a web-based cube puzzle game as a proof-of-concept prototype and conducted a user study. The basic gameplay is first introduced in section A, the selection of the visual concepts for this case study are motivated, and their implementation within the game is discussed in section B. Other visual concepts are reviewed, although intended for future exploration, in Section C. Finally, the actual images used in both versions of the game as well as their grouping to design the puzzles are detailed in Section D.

A. Baseline Gameplay

The Cube Puzzle game consists of a series of 6 puzzles, each with 9 pieces arranged in a 2D 3 3 layout (Fig. 3). Each piece represents a cube, and each of its six sides has a subimage of one of the six possible full images that can be formed with the 9 cubes. Only one of the six faces of each cube is displayed at a given time, and this face can be changed by rotating the cube in any of four directions (by clicking on the triangular marks) to reveal the corresponding neighbor face. In addition, the displayed face can be repeatedly rotated 90˝ clockwise by clicking on the corresponding inner arrows. Each puzzle is solved when all the cubes display the corresponding part of a target image and at the appropriate orientation, i.e. when the target image is formed. Although each cube has six faces, only one target image per puzzle has to be formed. Different puzzles have different images on the sides of the cubes. The images used for the six puzzles in this baseline version of the game are given in the first six rows in Table II. The rest of the information in this table is described where relevant in the following sections. A game is successfully complete when all the 6 puzzles have been solved. Note that this number (6) of puzzles has to do with the number of visual concepts we decided to work with, not with the number of sides of a cube.

B. Visual Concepts Considered

The baseline gameplay was enriched with visual computing (VC) elements (or visual concepts, for short) so that we can test our main hypothesis that this kind of visual concepts can be useful to modulate the gameplay. Much is known on the impact on human visual

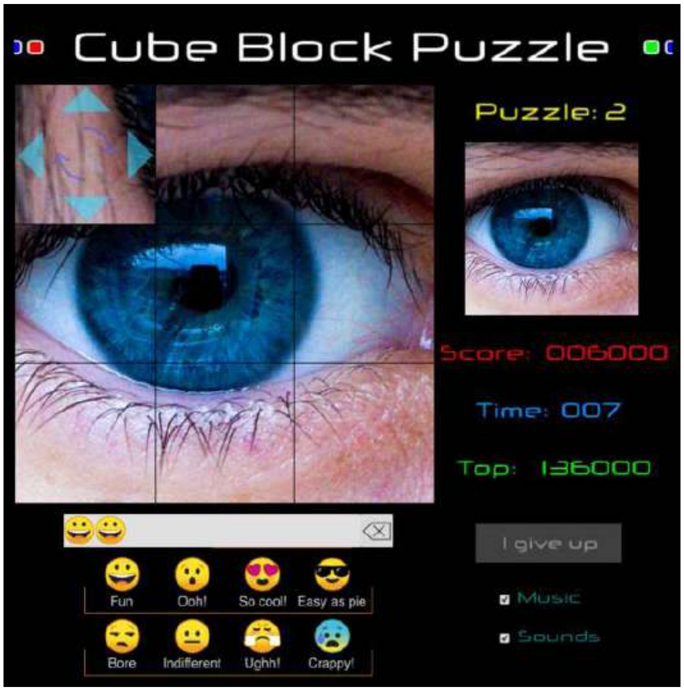


Fig. 3. Game interface with one of the puzzles. In this case, just a single cube (the one in the upper-left corner) remains to be aligned to complete the puzzle. The target image is a human eye in this case, which is displayed as a thumbnail for reference while playing.

perception of computer-generated graphics [26], even though a lot of practical insight has yet to be gained by the application of this body of knowledge to games, and to human-computer interfaces at large. For this first study, we decided to focus on the following three visual concepts, which are highly relevant to, and cover, three different areas of human visual perception: edges (*e*), color (*c*), and dynamics (*d*).

Edges can roughly represent a given object, but they can also be insufficient for object recognition [61]. Therefore, edges may challenge some visual-perception-based task.

Color is another visual attribute whose perception has also potential implications for games. For instance, color perception is known to differ across the visual field of view [62], and it is influenced by 3D shape perception [63]. Interestingly, brain activation from actual observations of color are found to be similar to *implied* observations, i.e. not actually observed, but known from prior knowledge of objects [64].

Motion is a powerful visual cue that can enhance the perception, particularly when other visual conditions are poor [65]. For our purposes, motion and image dynamics in general provide a wide range of concepts where time-varying contents can be used to modulate the discriminability of images. For instance, visual acuity of static stimuli is found to be superior than that of dynamic stimuli in the fovea, but not in the periphery [66], and perception is influenced by temporal frequency and occlusion [67].

Implementation details. Regarding the implementation of these concepts, for edges (puzzles $\mathsf{A}_{\mathrm{vc}:e}$ and $\mathsf{A}_{\widetilde{\mathrm{vc}}:e}$ in Table II), spatial image gradients were computed on the corresponding gray-level image, and then scaled for visualization purposes. Algorithm 1 shows the steps used in this work to convert a color image to an edge image. First, the image is converted to gray-scale by using the NTSC standard where each color channel is weighted differently ($w_r = 0.30$, $w_g = 0.59$, $w_b = 0.11$) to account for human color perception, which is more sensitive to green [25]. In order to reduce the noise effect, the gray-scale image is smoothed by average filtering [25], that is, a linear convolution, $I_g * A$, of the gray-scale image I_g with the filter mask A

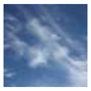
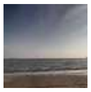
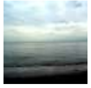
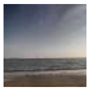
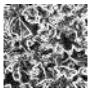
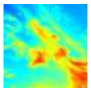
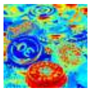
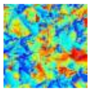
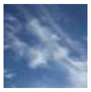
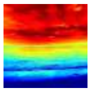
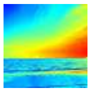
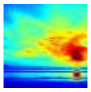
of size $n \times n$, with $A_{ij} = 1/n^2$. Spatial gradients [25] are computed from the smoothed image as $[G_x, G_y] = \nabla I_s = [\partial I_s / \partial x, \partial I_s / \partial y]$. The gradient magnitude is approximated by the sum of the absolute values of the gradient in both directions. We finally apply a scale factor s to the gradient magnitude for visualization purposes, and clamp the result to the valid gray-level range [0,255]. In this work, we used $n = 11$ for smoothing; spatial gradients were approximated with simple neighbor differences,

$$G_x(x,y) = \frac{\partial I_s}{\partial x}(x,y) \quad \approx \frac{I_s(x+1,y) - I_s(x-1,y)}{2} \tag{1}$$

$$G_y(x,y) = \frac{\partial I_s}{\partial y}(x,y) \quad \approx \frac{I_s(x,y+1) - I_s(x,y-1)}{2} \tag{2}$$

and $s = 30$ for scaling to visually emphasize the edges.

TABLE II. Images Used in the Different Puzzles for Each of the 6 Faces of Each Cube, I_i, $i \in \{0; ...; 5\}$. Each Image Is Divided Into 3 × 3 Images in a Regular Grid. These Subimages Are the Faces of Each Cube in the Puzzle (As Shown in Fig. 1 and Fig. 3). The Goal Is to Rotate Individually the Cubes so That the Visible Face of the Cube Correspond to the Respective Subimage in I0, Thus forming the Target Image

Puzzle	Target image (I_0)	Rest of images				
		I_1	I_2	I_3	I_4	I_5
A_{st}						
B_{st}				Same as A_{st}		
C_{st}				Same as A_{st}		
$A_{\widetilde{st}}$						
$B_{\widetilde{st}}$						
$C_{\widetilde{st}}$						
$A_{vc:e}$						
$B_{vc:c}$						
$C_{vc:d}$						
$A_{\widetilde{vc:e}}$						
$B_{\widetilde{vc:c}}$						
$C_{\widetilde{vc:d}}$						

Algorithm 1. Edge computation

Input: Color image with color channels R, G, B

Output: Edge map as image

```
$I_g \leftarrow w_r \cdot R + w_g \cdot G + w_b \cdot B$    ▷ convert to gray-scale image
$I_s \leftarrow$ averageFilter $(I_g, n)$    ▷ smooth image, with filter size $n$
$G_x, G_y \leftarrow$ spatialGradients $(I_s)$    ▷ vertical & horiz. image gradients
$M \leftarrow |G_x| + |G_y|$    ▷ approximation to gradient magnitude
return min $(s \cdot M, 255)$    ▷ scale and clamps to range [0, 255]
```

For color, a color map ("false color") was applied (puzzles $B_{vc:c}$ and $B_{\widetilde{vc:c}}$ in Table II). In this work, we use the known JET color map that we apply after converting the original image to gray-level scale (Algorithm 2). This color map and pseudo-color processing in general, are common aspects in image processing [25]. Color maps are functions (maps) from a scalar value (an index) in a range (typically gray levels in the range [0,255]) to tuples defining colors (e.g. RGB values). The JET color map (Fig. 4) produces a smooth transition from cold to warm colors as the index varies from 0 (the darkest) to 255 (the lightest).

Algorithm 2. Color modification

Input: Color image with color channels R, G, B

Output: Image in JET color map

```
$I_g \leftarrow w_r \cdot R + w_g \cdot G + w_b \cdot B$    ▷ convert to gray-scale image
return applyColorMap($I_g$, 'JET')    ▷ apply JET color map
```

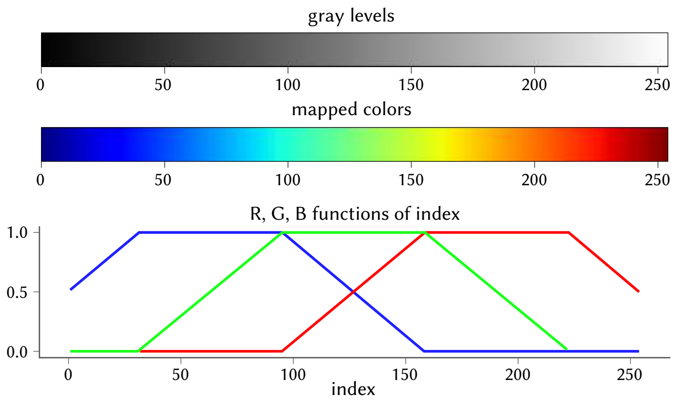


Fig. 4. JET color map. When applied to gray values (top) the corresponding colors (middle) are defined by specific functions (bottom).

Finally, the dynamics concept was implemented by clockwise rotating independently each of the six images around their center at a constant angular speed of about $\omega = 20°/s$. Therefore, when one subimage in one of the cube's sides is visible, the corresponding rotating motion is displayed, rather than a static image, as illustrated in Fig. 5. Such effect is applied to all images in puzzles $C_{vc:d}$ and $C_{\widetilde{vc:d}}$ in Table II. Note this transformation (Algorithm 3) is computed in real-time on a graphic processing unit (GPU), by means of the fragment shader [68]. To this end, images are provided as textures and faces are provided with 2D texture coordinates. So, for each frame, texture coordinates are transformed by the 2D rotating transformation matrix (3):

$$R(\varphi) = \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix} \tag{3}$$

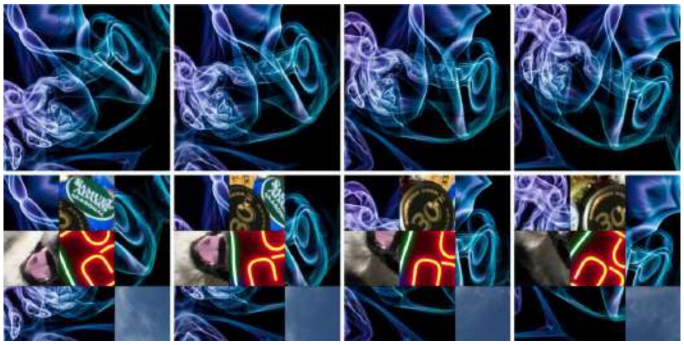


Fig. 5. Illustration of the dynamic effect. All images are rotating, including the target image (above), but only the parts corresponding to the disclosed cube side are visible at a given moment (below). This animation corresponds to 4 seconds, one frame per second, and the target image is visible in five out of the nine cubes.

```
Algorithm 3. Updating rotation effect, at each frame
Input: angular speed ($\omega$, degrees/second), and elapsed time ($t$, seconds)
$\varphi \leftarrow$ rotationAngle $(\omega, t)$          ▷ update rotation angle
$R \leftarrow$ rotationMatrix $(\varphi)$        ▷ create 2D clockwise rotation matrix
Send R to fragment shader    ▷ for rotation of all textures coordinates
```

For the experiment, all the color images were of 512 × 512 pixels, and the visual modifications corresponding to edges and color change were precomputed and used in the video game afterwards, while the dynamic effect was performed in real-time.

C. Other Visual Concepts

Besides these three concepts, we comment on a few other interesting possibilities that might be explored in the future. Due to its fundamental importance in our daily lives, human faces have been extensively studied from both computer and human perception points of view [69]–[71], and might be a good concept to include and study in the context of digital games; for instance, face-based environments may increase users' engagement [72].

Pixelation is a simple but powerful mechanism to modify images so that their perception can be facilitated or hindered [73], and it lends itself to be included in video games for a variety of purposes. However, pixelation in successful games has only or mostly been used for its nostalgic or art style values [74], [75].

Visual illusions are an important area of vision research since they provide great insights into the functioning, abilities and limitations of our visual system [76]. Interestingly, it was recently found that machines cannot (yet) understand nor create such illusions [77]. Consequently, illusions might arguably be another ingredient in video games, a topic worth exploring in the future.

D. Target Images

Regarding the target images, three generic visual entities are considered: human (facial) attributes, natural landscapes, and abstract shapes. The particular images chosen for representing these entities were eyes, beaches, and smokes, respectively (column I_0 in Table II). One main reason for these images is that each of them lend themselves, respectively, to each of the visual concepts chosen. Thus, edges were applied to eye images, color modification to the beach scenes, and dynamics to the smoke images.

Additionally, for each puzzle, two degrees of likeness (`distinct` and `alike`) are considered depending on which images are used for each of the remaining five faces (columns I_1–I_5 in Table II) of the cubes other than the face for the target image. For the `distinct` version, images very different from the target ones are used, whereas the `alike` version uses distracting images that look similar to the target image. For instance, for the eyes case, other human eyes with similar appearance are used as well.

Summarizing, there are two versions of the game: ST (after "standard") as the original one, without visual computing; and VC (after "visual computing"), which uses the three visual concepts (edge, color and dynamics), one per target image. For each version of the game (ST and VC), two factors are considered: the three target images (eyes, beaches and smoke), and the type of distracting images (distinct and alike). In a way, using distracting images might be seen as an additional VC to consider, since it may affect the gameplay as well. Thus, this hypothesis will also be tested experimentally.

Table III collects all this information for convenience and quick reference. We will refer to each of the three puzzle pairs by different target images (A, B, C). The alike versions are denoted as $A^{\sim}_{i}$, $B^{\sim}_{i}$, and $C^{\sim}_{i}$ (for the corresponding $i \in \{st, vc\}$). This notation will allow us to compactly and easily refer to the different puzzles, in particular when comparing their results. As an example, the puzzle shown in Fig. 3 corresponds to $A^{\sim}_{st}$, and that shown in Fig. 5 corresponds to $C_{vc:d}$ as images are rotating. All images used in all puzzles are given in Table II. These are all static images except for puzzles $C_{vc:d}$ and $C^{\sim}_{vc:d}$, which move as indicated.

TABLE III. The Conditions and Notation to Refer to the 12 Different Puzzles Used, 6 Puzzles Per Game Condition. TABLE 2 Provides Useful Complementary Information to Understand Each Possible Puzzle

Puzzle	Target image	Other 5 images	Likeness	Visual concept	Condition	
					ST	VC
A	eye	not eyes	distinct	edges (e)	A_{st}	$A_{vc:e}$
		other eyes	alike (~)		$A^{\sim}_{st}$	$A^{\sim}_{vc:e}$
B	beach	not beaches	distinct	color (c)	B_{st}	$B_{vc:c}$
		other beaches	alike (~)		$B^{\sim}_{st}$	$B^{\sim}_{vc:c}$
C	smoke	not smokes	distinct	dynamics (d)	C_{st}	$C_{vc:d}$
		other smokes	alike (~)		$C^{\sim}_{st}$	$C^{\sim}_{vc:d}$

IV. Design and Implementation

We review the rationale behind the design of the experiment in section A, the user questionnaire after game completion in section B, and the interface in section C. Details of the user study and of the implementation are given in section D and section E, respectively.

A. Experiment Design

Two user groups are considered, the control group (ST) and the experimental group (VC). A between-subjects protocol was preferred over within-subjects one, since less time and effort is required from each participant, and the subjects do not need to compare directly both versions of the game. Although the between-subjects protocol requires more subjects, this was not found to be an issue, because we planned to do an online experiment and expected to recruit participants with relative ease. Interaction logs were saved per session in order to collect quantitative data in terms of time to completion, number of puzzles successfully solved, number of clicks required, etc. Additionally, subjective data was captured via a final opinion questionnaire (Sect. B) to understand how easy, entertaining, or enjoyable the game was perceived by players. Qualitative feedback from users data was also gathered during and at the end of each puzzle (Sect. C). Table IV summarizes the main descriptions of the experimental design.

TABLE IV. Summary of Main Descriptions of the Experimental Design

Study type	Online game user study for about 1,5 months. Each participant played 6 puzzles for up to about 15 minutes (Sect. IV-D).
Experimental design	Between-subjects (Sect. IV-A): • Control group (ST): original (standard) images. • Experimental group (VC): images with visual modifications.
Assignment to experimental condition	Uniformly at random (either ST or VC)
Statistical hypothesis testing	Mann–Whitney U test (Fig. 8), Kruskal-Wallis H-test (Fig. 8), χ^2 test (Sect. 5-C-1), Z score test (Tables XII and XIII)
Exploratory analysis	Confidence intervals (Fig. 6) and effect sizes, Cohen's *d* (Fig. 7)
Intervention	Visual modifications and experimental conditions (Table II and Table III)

Online participants were assigned randomly either to the ST or the VC conditions, but they were not given any information about the game conditions so that no expectation could bias their judgment and play, a danger that has been identified [60]. All users are requested to complete 6 puzzles, but they can cancel a particular puzzle if they prefer so (Sect. C). To avoid a presentation-order effect, the three pairs of puzzles were randomly presented, but within each pair, the `distinct` puzzle was always presented before the `alike` one. For instance, one possible presentation order in the control group might be B_{st}, $B_{\widetilde{st}}$, C_{st}, $C_{\widetilde{st}}$, A_{st}, $A_{\widetilde{st}}$, and a possible presentation order in the experimental group might be $A_{vc:e}$, $A_{\widetilde{vc:e}}$, $B_{vc:c}$, $B_{\widetilde{vc:c}}$, $C_{vc:d}$, $C_{\widetilde{vc:d}}$. Thus, a total of $3! = 6$ different presentation orders are possible.

To avoid an effect due to the type or contents of the images, the same images were given to all participants, except obviously for the corresponding VC modifications in the control group. Since it is important to find out the completion times, the game is timed and the player is asked to proceed as quickly as possible. To prevent interruptions from being included in the total elapsed time, the player is offered to take breaks after each puzzle, if desired, and this time is not considered. The elapsed time is displayed while forming a puzzle (Fig. 3) to provide the user with feedback and awareness of the time.

B. Opinion Questionnaire

The contents of the questionnaire (Table V) were designed to learn about how much the player enjoyed the game as a whole (Questions 2 and 4), and their subjective perception of elapsed time (Question 3) and performance (Question 1). We were also interested in finding out whether they enjoyed or hated some particular puzzle (Questions 13–14). To better understand in terms of what they liked or disliked the game, their overall experience was assessed in four dimensions (Questions 5–8). A drawback of the between-subject approach is that, since each participant only plays one version of the game, they cannot be asked which one they prefer. To cope with this, users were given a number of widely known reference games to compare with (Questions 9–12), so that the two versions of the game could eventually be comparable indirectly via these reference games.

The parts of the questionnaire and the possible answers in each question were as follows:

- Questions 1–4 could be answered with a 5-level Likert scale (from "strongly disagree" to "strongly agree"), plus a "No answer" choice.
- Questions 5–8 inquired about particular dimensions of the overall experience, and the possible answers were as given (Table V), plus a "No answer" choice.

TABLE V. Questionnaire Contents

No.	Question
1	In general, I found easy to complete the game
2	I found the game entertaining
3	I think I was quick in solving the puzzles
4	Would you play this game again?
5-8	I found the overall experience to be...
5	... entertaining | boring
6	... simple | complex
7	... surprising | dull
8	... exciting | frustrating
9-12	I liked this game (significantly more | more | similarly to | less | significantly less) than
9	... Mahjong
10	... Solitaire
11	... Classical puzzle
12	... Sliding puzzle
13	Which puzzle did you like the most?
14	Which puzzle did you like the least?

- Questions 9–12 seeks to find out how much the player likes this particular game in comparison to the reference games given. A representative image of each reference game is given to help the player recognize the game.
- Questions 13–14 are about which puzzle they liked the most or the least, and they are offered to choose among the six puzzles. To facilitate their recognition, both the target and the remaining images are presented.

Notice that Questions 1 and 3 relate to the *perceived* effort, and the main motivation for including them is to relate them with the measured, actual effort in terms of times and number of clicks required to complete the puzzles. Our intuition was that a lack of correlation between the perceived and the actual efforts might provide some clue on whether the subjects enjoyed the game. For instance, if they underestimate their effort, this might mean they have been losing track of time, which is a sign of engagement [1], without explicitly asking the participant on that.

C. Interface Design

Some interface design details that are particularly important for this study are now briefly discussed. Before starting with the actual puzzles, the player is offered to play a first puzzle with neutral images so as to make sure they understand how to play. We will be referring to this initial "learning" puzzle as `L`.

It turns out that it can be very frustrating for a player not to be able to complete one puzzle that can be found particularly challenging, since this would prevent them to continue with the remaining puzzles. This is addressed with a "Give up" button (Fig. 3). However, to help preventing a player from giving up too easily, this button only becomes active after two minutes from the starting time of each puzzle. As a form of feedback, time is displayed as a countdown in seconds, starting with 120. When it gets to zero, the game is still playable and the player can give up.

The scoring captures the performance in terms of completing each puzzle and parts of it, and the time elapsed (Table VI). The first test puzzle is not included in the scoring. This score was included to provide the player with a sense of an actual game, rather than as a reliable measure of performance to be analyzed. When the sides for all the 3 cubes in a row or in a column are correct, the player is provided with feedback, which is also a form of visual reward and encouragement to keep playing.

TABLE VI. Scoring Scheme, Points are in Thousands

Situation	Points
Completing one row or column	1
Completing one row and column at the same time	4
Puzzle solved	10
For each remaining second once puzzle completed	2

Although the user is asked to fill in a final questionnaire, this only provides us with a form of *overall* opinion, but it was interesting to get more detailed feedback from the user for each of the puzzles, so that the player could express their emotional feelings at a time closer to the moment they experience them. To that end, an Instant Emotional Feedback (IEF) component was designed in the form of emojis (bottom part in Fig. 3). This IEF allows the user to choose, at any time during a puzzle or right before moving to the next puzzle, one or more among 8 possible emojis to express a subset of the emotions most closely resembling their mood.[1]

An individual emoji can be selected more than once as a form of indication of the strength of the corresponding emotion. Although these feelings can be roughly categorized into positive and negative emotions (Table VII), we used this just as a pragmatic approximation for usability, since it offers a convenient grouping criterion. For instance, a surprise is not necessarily a positive emotion (it could be a surprise for bad); and the expressions for difficulty can convey a negative valence if a puzzle is felt as *too much* difficult, as this would depart from the ideal flow concept. So, the interpretation of the use of some of these emojis should be taken as a mere approximation of the true feelings.

TABLE VII. Emotions and Emojis for the IEF (Instant Emotional Feedback). Since the Emoji Images Can be Ambiguous in the Associated Feeling, a Short Caption was Added Below

Valence	Underlying emotion, emoji used to convey the emotion, and informal caption			
positive	Entertaining 😀 Fun	Surprising 😯 Ooh!	Exciting 😍 So cool!	Easy 😎 Easy as pie
negative	Boring 😒 Bore	Dull 😐 Indifferent	Frustrating 😤 Ughh!	Difficult 😰 Crappy!

D. Pilot and Final User Study

After testing the game ourselves, four people of different profiles (two of them females; one senior lecturer, one teenager high-school student, and two young students, one undergraduate and one graduate), were asked to play the game to identify potential functionality and usability issues. Each game condition (ST and VC) were assigned to two of them. No problem was identified, and all of them could complete the game. Interestingly, two of the pilot users reported to find the initial puzzle critically important to understand that only one image needed to be formed, and which one it was. The feedback after a row or column is complete was reported to be either useful (to make sure whether a puzzle is fully solved) or satisfying.

For the final study, calls for participation were submitted to mailing lists in our university, both to staff (lecturers and administration) and students, both from different disciplines (humanities, law and economics, health, and sciences & engineering), so different ages and backgrounds were covered. Users were informed on the scientific purpose of the study and that the privacy of the collected data was guaranteed since it would only be treated anonymously.

For the statistical significance analyses, the following tests were used: the non-parametric Mann–Whitney U test (M-W for short), Kruskal-Wallis H-test, χ^2 test, and Z score test, depending on the nature of the data and the comparison purpose. Even though we report p-values, we would like to emphasize the danger of dichotomous thinking associating statistical significance to conclusive evidence (and non-significance to no evidence) [80], that statistical testing may promote. Related to this, and also because our work can be seen as a mixture of confirmatory and exploratory research [81], we complement the significance testing with an estimation-based approach based on confidence intervals and effect sizes.

E. Implementation Details

All parts of the game, including the questionnaire, were implemented using HTML5 and JavaScript, and WebGL for the graphics. Since users did not register for playing and no personal information from them was collected, the logged actions, the results of the different puzzles, and the responses to the questionnaire were associated with randomly generated names. The game condition (ST or VC) was selected uniformly at random, so that approximately half of the participants were assigned to each group.

The game was (and is) made available at https://bit.ly/3dxLFXZ. The link was provided as part of the message for the participation, for convenient quick and ready access. This is the version used during the study, where the game condition is chosen randomly. For the reader convenience, VC and ST versions can be accessed at https://bit.ly/3lRLV8z and https://bit.ly/ 3IyaNfo, respectively.

V. Results and Analysis

Up to 271 users started the game, and 55% of them (148) completed it. According to the random assignment to the control (ST) and experimental (VC) groups, about half of the participants played under each condition (Table VIII). Interestingly, a higher percentage of ST participants compared to VC participants completed the game, which can be seen as a first sign that the VC version of the game might be more challenging, since it may have caused some users to quit at some point before finishing.

TABLE VIII. Number of Participants who Completed the Game

Condition	Started	Completed (%)
ST	126	77 (61.1)
VC	145	71 (49.0)
Total	271	148 (54.6)

We now discuss the players' performance in section A. Then, the instant emotional feedback and the answers to the questionnaire are analyzed in sections B and C, respectively. All these analyses focus on the data from the users who completed the game, to avoid the bias that considering the data from all participants may introduce.

A. Behavioral Results

In order to study how challenging the different puzzles and game conditions were, two play performance metrics are mainly considered: the time taken to complete the game, and the number of clicks (i.e. cube rotations) the users made. Higher values for these metrics can readily be associated to an overall greater difficulty. We analyze these metrics globally in section 1 and per-puzzle in section 2.

[1] Notice that this form of IEF might be related to the different approaches of getting feedback by monitoring player's physiological signals [78], [79], which are more costly and obtrusive, but have the potential to predict the players' affective state in real time.

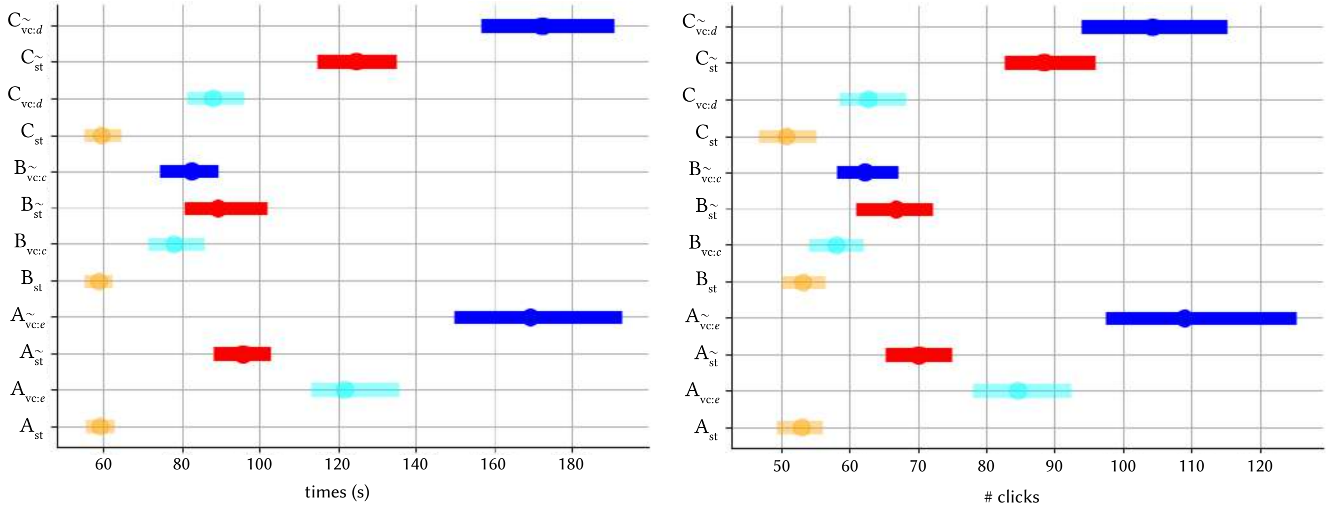


Fig. 6. Times (left) and clicks (right) per puzzle, with 95% confidence intervals.

B. Global Analysis

On average, VC users took 50% longer and required about 25% more clicks than the ST users for game completion (Table IX). These differences are statistically significant (U = 763.0 for times and U = 1292.5 for clicks, p < 0.0001 both cases), and provide strong evidence on the higher challenge that visual computing introduces. This first finding leads to two relevant questions: which particular visual concepts are offering higher challenge, and how participants subjectively perceive the game. These issues are analyzed in section 2 and section C, respectively.

TABLE IX. Performance Metrics of Game Completion: Mean (Standard Deviation)

Metric	All users	ST	VC
Time (Seconds)	594 (189)	486 (115)	710 (183)
# clicks	429 (106)	384 (68)	479 (117)

Although somehow more anecdotally, 19 users used the "give up" button for particular puzzles. Most of of these users (14) were playing the VC version, which brings further evidence on previous observations. Curiously enough, these 19 users took about 20% longer to complete the game than the remaining 129 users. Individually, any of those players who gave up some puzzle, used this option twice at most. These observations suggest that the give-up option was used sparingly, and only when players sensibly judged they were taking too long to complete some particular puzzle. Most surrenders correspond to puzzle $C_{\widetilde{vc:d}}$ followed by $C_{vc:d}$, a fact that can be further inspected in light of other results discussed in subsequent sections.

C. Per-Puzzle Analysis

The above overall results can be broken down by the different kind of puzzles. The confidence intervals [82], [83] for time and clicks (Fig. 6), computed with bootstrapping (100 repetitions, percentile method [84]), provide both a good first idea of which puzzles require more effort to complete, and qualitative support to the quantitative analysis described subsequently.

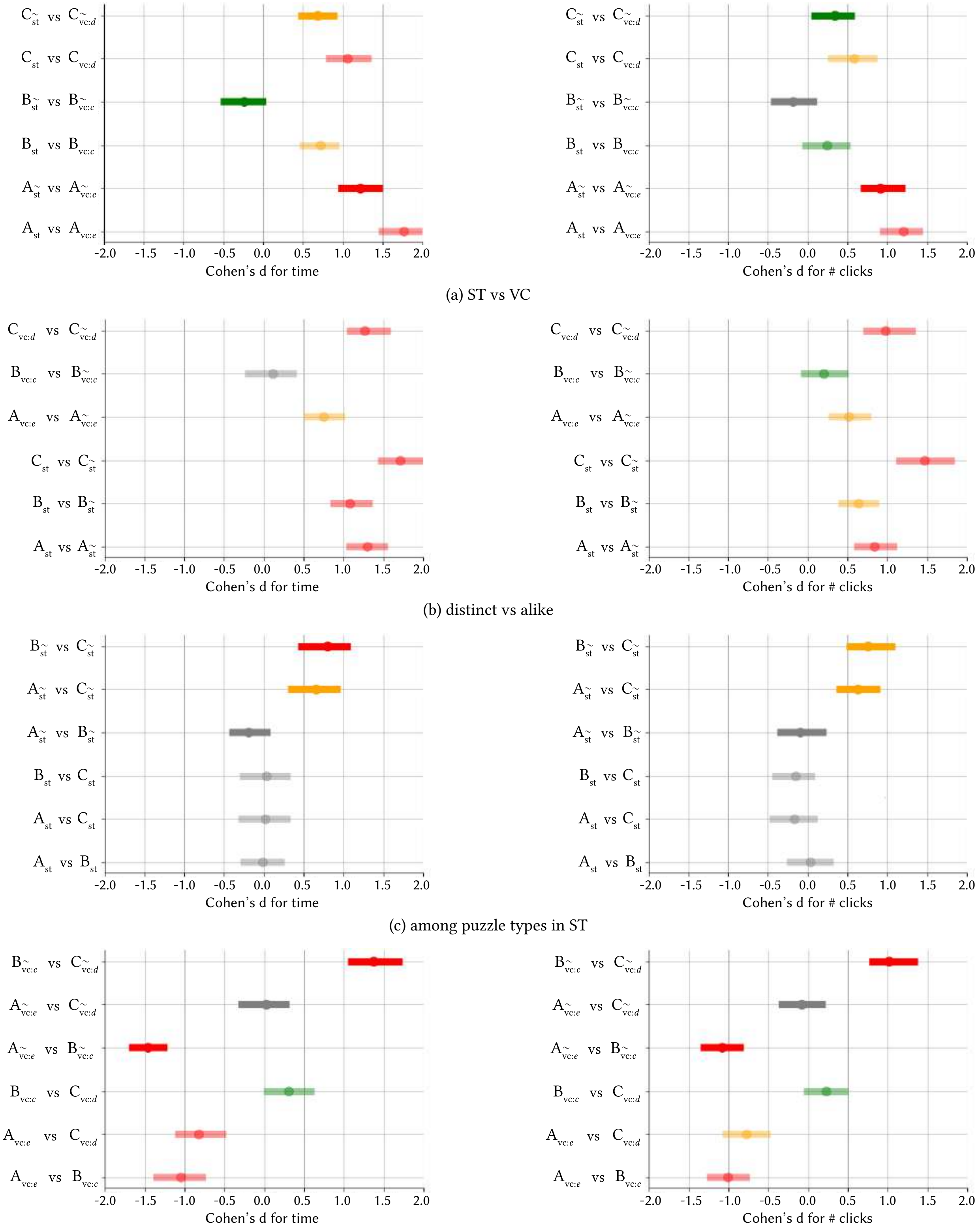
Three comparisons are considered: ST versus VC, `distinct` versus `alike` puzzles, and among puzzle types (namely, with different target images for the same game condition and likeness). These analyses are performed in two complementary manners: effect sizes and hypotheses testing, as reasoned above (Sect. D).

For the effect sizes, Cohen's d [83] was used (Fig. 7(a–d)) where the confidence intervals for the mean were computed with bootstrapping as well. For convenience, the intervals were color-coded according to the known effect size classes (small, medium and large) [83] using the mean as a reference, as follows: large if $|d| \geq 0.8$, medium if $0.5 \leq d < 0.8$, small if $0.2 \leq |d| < 0.5$, and none if $|d| < 0.2$. For the statistical tests, all the comparisons are given graphically (Fig. 8a–c) for higher clarity.

ST versus VC. Players' performance for each of the 6 puzzles plus the initial puzzle (`L`) were pair-wise compared under the ST and VC conditions (Fig. 7a, Fig. 8a). Significant differences were found (Fig. 8a) for both the required completion time and number of clicks, for all puzzles except for `L` and `B˜`. These results are consistent with those obtained with Kolmogorov-Smirnov test, with the only difference being in $C_{\tilde{st}}$ vs $C_{\widetilde{vc:d}}$ for times (not for clicks); thus both tests agree in 11 out of 12 comparisons. The statistical differences are also essentially captured with the effect sizes (Fig. 7a). The lack of differences in `L` ($U = 2617.5$, $p = 0.33$ for time, and $U = 2700.0$, $p = 0.45$ for the number of clicks) makes sense because this is just a test puzzle not including any form of visual computing. Thus, this result serves as a verification that no undesirable bias due to differences in subject distribution exists. As for the lack of differences between $B_{\tilde{st}}$ and $B_{\widetilde{vc:c}}$ ($U = 2408.5$, $p = 0.11$ for time, and $U = 2454.0$, $p = 0.14$ for the number of clicks), it might relate to the fact that the alike version in this puzzle (i.e. with distracting images similar to the target image) had a higher impact on the difficulty of $B_{\widetilde{vc:c}}$ than the fact of including the VC. This suggests that the distracting images can be regarded as a form of visual computing itself, as initially hypothesized.

Distinct versus alike versions. Performances were also compared between the distinct and alike versions of the puzzles for both the VC and ST conditions, separately, i.e. A_{st} vs $A_{\tilde{st}}$, $A_{vc:e}$ vs $A_{\widetilde{vc:e}}$ and so on (Fig. 7b, Fig. 8b). With a confidence level $\alpha = 0.05$, significant differences are found (Fig. 8b) in both ST and VC, for the three types of puzzles, except for $B_{vc:c}$ vs $B_{\widetilde{vc:c}}$, a result which agrees with the per-puzzle analysis described above. It can be observed that the medium-big effect sizes (Fig. 7b) correspond to the statistical differences. Notice that the differences found in the ST condition further support the usage of distracting images as a factor for game-difficulty modulation.

Among puzzle types. Finally, to gain insight on whether some puzzles may be harder than others, performances were compared group-wise for two 3-puzzle groups, namely, the three `distinct` puzzles, and the three `alike` ones, again for VC and ST separately (Fig. 7c–d, Fig. 8c). The differences among the puzzles of the distinct group for the ST condition (A_{st}, B_{st}, C_{st}) seem statistically insignificant, which agrees with Cohen's d close to 0, which means that none of the different images being used (eyes, beaches and smoke) bring any particular challenge with



(a) ST vs VC

(b) distinct vs alike

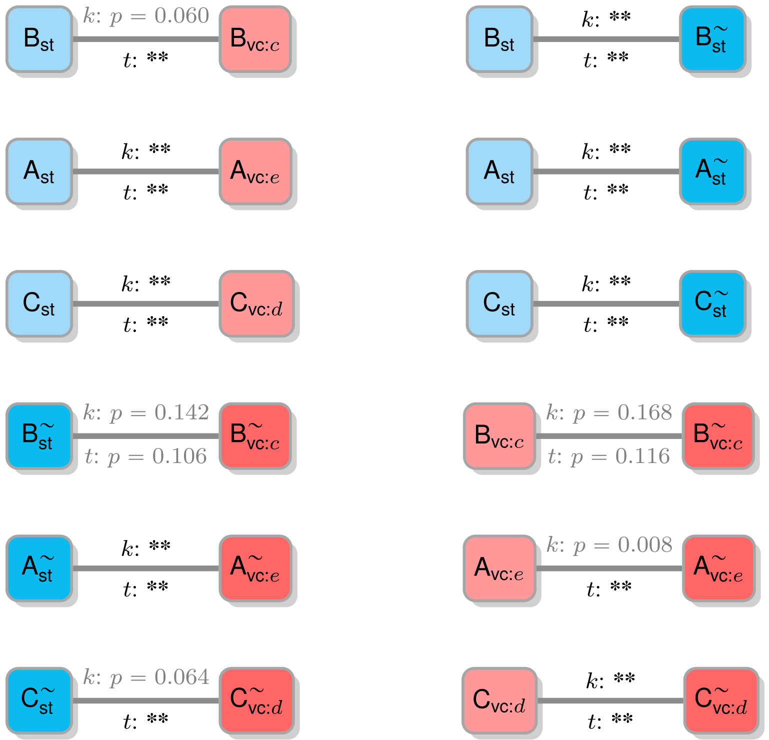
(c) among puzzle types in ST

(d) among puzzle types in VC

Legend for interval's color:

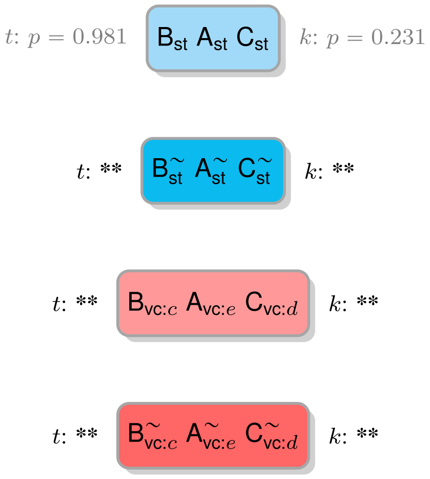
Effect size		distinct vs alike (in a, c, d)
red	big	light distinct
yellow	medium	**dark** alike
green	small	
gray	none	

Fig. 7. Confidence intervals of effect sizes (cohen's *d*) for times (left) and clicks (right) when comparing different puzzles.



(a) Pairwise ST vs VC

(b) Pairwise distinct vs alike


t: $p = 0.981$
B_{st} A_{st} C_{st}
k: $p = 0.231$
t: **
$B_{\widetilde{st}}$ $A_{\widetilde{st}}$ $C_{\widetilde{st}}$
k: **
t: **
$B_{vc:c}$ $A_{vc:e}$ $C_{vc:d}$
k: **
t: **
$B_{\widetilde{vc:c}}$ $A_{\widetilde{vc:e}}$ $C_{\widetilde{vc:d}}$
k: **

(c) Group-wise within distinct and alike

(i) ST

Group	Rank	
	Time	# clicks
distinct	—	—
alike	$A_{\widetilde{st}} \prec C_{\widetilde{st}}$ $B_{\widetilde{st}} \prec C_{\widetilde{st}}$	$A_{\widetilde{st}} \prec C_{\widetilde{st}}$ $B_{\widetilde{st}} \prec C_{\widetilde{st}}$

(ii) VC

Group	Rank	
	Time	# clicks
distinct	$B_{vc:c} \prec A_{vc:e}$ $C_{vc:d} \prec A_{vc:e}$	$B_{vc:c} \prec A_{vc:e}$ $C_{vc:d} \prec A_{vc:e}$
alike	$B_{\widetilde{vc:c}} \prec A_{\widetilde{vc:e}}$ $B_{\widetilde{vc:c}} \prec C_{\widetilde{vc:d}}$	$B_{\widetilde{vc:c}} \prec A_{\widetilde{vc:e}}$ $B_{\widetilde{vc:c}} \prec C_{\widetilde{vc:d}}$

(d) Post-hoc test after (c)

Fig. 8. Diagrams of the statistical tests performed on comparing types of puzzles (a, b) pairwise (Mann-Whitney rank test) and (c) group-wise (Kruskal-Wallis H-test), for completion time (t) and number of clicks (k). Bonferroni correction is applied with $\alpha^* = 0.05/m$, for $m = 48$ accounting for the all pair-wise hypotheses (24) and the two dependent variables (t and k). Note: ** means p-value is very low (lower than 0.0001), thus representing statistically significant differences (d) Ranking puzzles from pair-wise post-hoc tests (Mann-Whitney rank test). Note: the notation X < Y represents that significant differences exists between X and Y and puzzle X has lower mean for the corresponding metric than puzzle Y.

respect to the others. However, when either VC concepts ($A_{vc:e}$, $B_{vc:c}$, $C_{vc:d}$) or alike versions ($A_{\widetilde{st}}$, $B_{\widetilde{st}}$, $C_{\widetilde{st}}$), or both ($A_{\widetilde{vc:e}}$, $B_{\widetilde{vc:c}}$, $C_{\widetilde{vc:d}}$) are introduced, the remarkable differences in both metrics (time and the number of clicks) suggest that the visual computing elements induce differences that affect both the `distinct` and `alike` versions of the puzzles.

Now, to understand which puzzles are harder than others, the three possible pair-wise post-hoc tests are performed for the cases where group differences have been found. From the results (Fig. 8d) the most remarkable outcome is that $C_{\widetilde{st}}$ seems the most difficult one among the alike versions under ST. For VC, $A_{vc:e}$ seems the most challenging one among the distinct versions, and $B_{\widetilde{vc:c}}$ the easiest one among the alike versions. Interestingly, both the times and the number of clicks agree in the puzzle ranking. This suggests that both metrics characterize similarly the control and experimental groups. Regarding the effect sizes, the medium to big sizes found in both ST and VC (Fig. 7c-d) are in agreement with the statistical results. Effect sizes also reveal a close symmetry between times and clicks.

Taken together, these observations can be summarized as follows. First, the game difficulty can be modulated not only by visual computing but also by a proper choice of the target and distracting images. Second, interestingly, the VC elements bring an additional difficulty beyond image contents *even* with the `distinct` versions of the puzzles, which provides an evidence of its intended effectiveness. Finally, remarkable differences exist among the three types of VC elements introduced, which promises to offer flexibility when choosing the desired challenge level for a target game and user profile.

D. Instant Emotional Feedback

Beyond the performance metrics, we are also interested in finding which additional insights can be gained from the qualitative instant emotional feedback (IEF). Around 90% of the users who completed the game used some emoji to provide their reaction to particular puzzles (Table X), in similar amounts for users in ST and VC groups, just slightly higher for the latter. We analyze which emojis were used globally in section 1, and per-puzzle in section 2.

TABLE X. Usage of the Instant Emotional Feedback (Emojis)

Condition	Users (%)	# Emoji	# Emojis/user (avg.)
ST	88.3	478	6.2
VC	92.9	517	7.3
Total	89.9	995	6.7

E. Global Analysis

It was found that the VC group completed the game with more time and clicks, which we can associate to a generally higher difficulty of the game under the VC condition. Now, based on the emojis chosen by the users (Table XI), it can be observed that emotion easy is used more than hard in the ST group, and the opposite happens in the VC group, which reinforces such association. Besides this increment in the perceived difficulty in the VC group, emotion `entertaining`, although decreasing, maintains the first place, while the rest of positive emotions (`surprising` and `exciting`) increase. In other words, by considering the group of the three emotions (`entertaining`, `surprising`, and `exciting`) as a whole, they are used similarly in ST (56.9%) and VC (53.4%), which can be interpreted that, overall, emotions with positive valence are similarly expressed in both conditions, which support our hypotheses regarding the game modulation and diversification through VC. Similarly, although hard is used more in VC than in ST, other emotions with negative valence (dull, frustrating and boring), not only are used the least in both conditions, but also with the same percentage (17.8% in total).

TABLE XI. Percentage of Emotions Usage (With Respect to the Total Emotions Used Group-Wise) in Both Groups, Sorted in Descending Order

ST		VC	
Emotion	%	Emotion	%
entertaining	31.8	entertaining	21.5
surprising	14.4	hard	19.1
easy	14.0	surprising	17.0
hard	11.3	exciting	14.9
exciting	10.7	easy	9.7
dull	10.0	dull	6.6
frustrating	6.1	frustrating	6.4
boring	1.7	boring	4.8

When comparing the usage proportion of each emotion (Table XII), it turns out that ST users found the game more entertaining, while VC users found it not only harder (which brings further statistical support to previous analysis) but also more boring and more exciting at the same time. Although this result clearly indicates that a per-puzzle analysis is called for, it is also worth noting that the result for `boring` is possibly less relevant, since this emotion is used little (less than 5%). Also, note that after Bonferroni correction, only the differences regarding `alike` are found significant.

TABLE XII. Emotions Whose Usage Proportions are Found Statistically Different (*Z* score, *A* = 0.05). The Symbol + Represents in Which Group (ST Or VC) the Corresponding Emotion is More Used, Whose % of Usage is Also Given

ST	VC	% usage	*p*-value	*z*
+ entertaining		31.8	0.046	1.997
	+ **hard**	19.1	**	-4.517
	+ exciting	14.9	0.011	-2.528
	+ boring	4.8	0.016	-2.404

Note: ** means *p*-value is very low (lower than 0.00001). After Bonferroni correction with $\alpha^* = \alpha/m$, for the m = 8 emotions tested, the difference in proportion is significantly different only for hard, which is boldfaced.

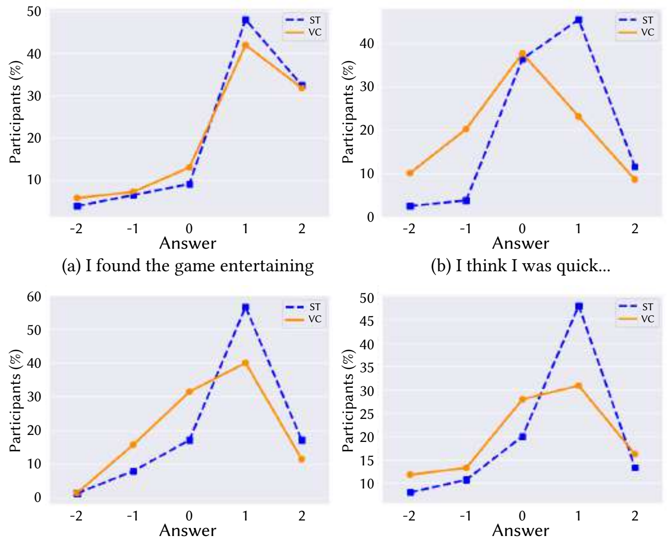
F. Per-puzzle Analysis

When the emoji usage is analyzed pair-wise (ST vs VC) per puzzle (Table XIII), it is found that significantly more emojis were used in VC than in ST for all puzzles: up to 9 emotions are found to be significantly more used in VC than in ST, while only 2 emotions were more used in ST, which additionally occurs in just two puzzles (A and A˜)[2]. This seems to imply that the VC condition generally rouses more emotions in a wider range of conditions. Another interesting observation is that although entertaining was found to be predominant in ST in a global sense (Table XII), its significance actually only occurs in a single puzzle ($A^{\sim}_{st}$), which suggests that it cannot be generalized the fact of ST being more entertaining than VC. A final remark is that, as noted above, no significant difference was found in the time required to solve B˜ between ST and VC (Fig. 8a). However, according to the IEF, B˜ is found easier in VC, which not only further supports the idea that distracting images produces higher difficulty, but also that VC modulates this difficulty and, very interestingly, by lowering it in this particular case. Notice that this analysis changes after Bonferroni correction.

TABLE XIII. Emotions Used Significantly More (*Z* score, α = 0.05) Between the ST and the VC Versions of Each Puzzle. After Bonferroni Correction With α* = α/*M*, for the *M* = 48 (6 Puzzles and 8 Emotions), the Difference in Proportion is Significantly Different Only for the Emotions in A, Which are Boldfaced

	A	*z*	*p*-value	A˜	*z*	*p*-value
ST	+ **easy**	+3.319	0.001	+ entertaining	+2.458	0.0140
VC	+ **hard**	-3.616	0.0003	+ hard	-2.619	0.0088
	+ **frustating**	-3.436	0.001			
	B			B˜		
VC	+surprinsing	-2.156	0.0311	+ easy	-2.000	0.0456
	C			C˜		
VC	+ hard	-2.438	0.0148	+ hard	-2.155	0.0312
	+ exciting	-2.000	0.0456	+ exciting	-3.000	0.0026

[2] We remind the reader that these are not necessarily the first puzzles presented to the player, since the order is chosen randomly for each participant, as discussed in A.



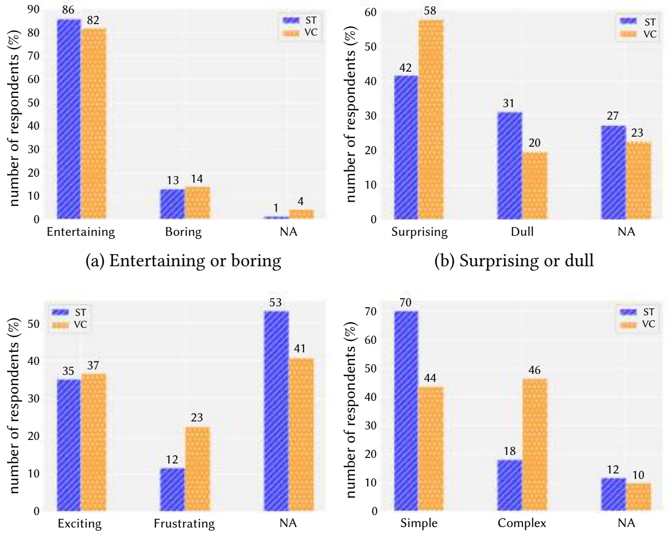
(a) I found the game entertaining

(b) I think I was quick...

(c) I found easy to complete...

(d) I would play the game again

Fig. 9. Distribution of replies to Questions 1–4 of the questionnaire. Please, note that a numerical scale has been used for the agreement scale, as given below, for convenience.



(a) Entertaining or boring

(b) Surprising or dull

(c) Exciting or frustrating

(d) Simple or complex

Fig. 10. Distribution of replies regarding the overall game experience. We include the percentage of users that did not select (NA) any of the adjectives of each pair.

G. Opinion Questionnaire

Finally, the participants opinions are analyzed mainly to find out how each game condition (ST or VC) was perceived (Sect. 1) and then to compare subjective perceptions with actual performance metrics (Sect. 2).

1. Participants' Perception

The responses of the questionnaire (Figs. 9–11) were compared pair-wise (ST vs VC) per question. No statistically significant differences were found in most cases. A general qualitative observation of the responses (Fig. 9) reveals that most users found the game entertaining (Fig. 9a), and easy to complete (Fig. 9c), and they agreed that they would play the game again (Fig. 9d). Most of the users in the ST group also agreed they were quick at solving the puzzles (Fig. 9b).

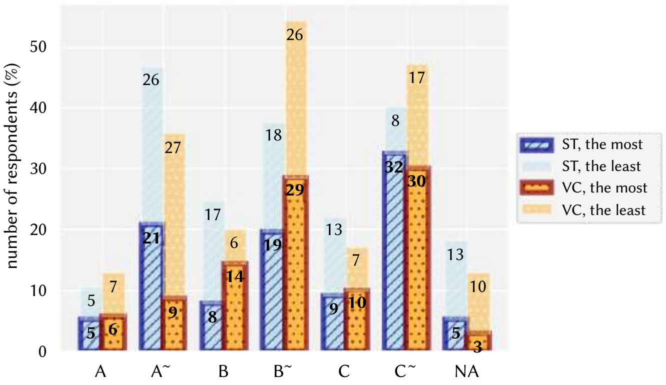
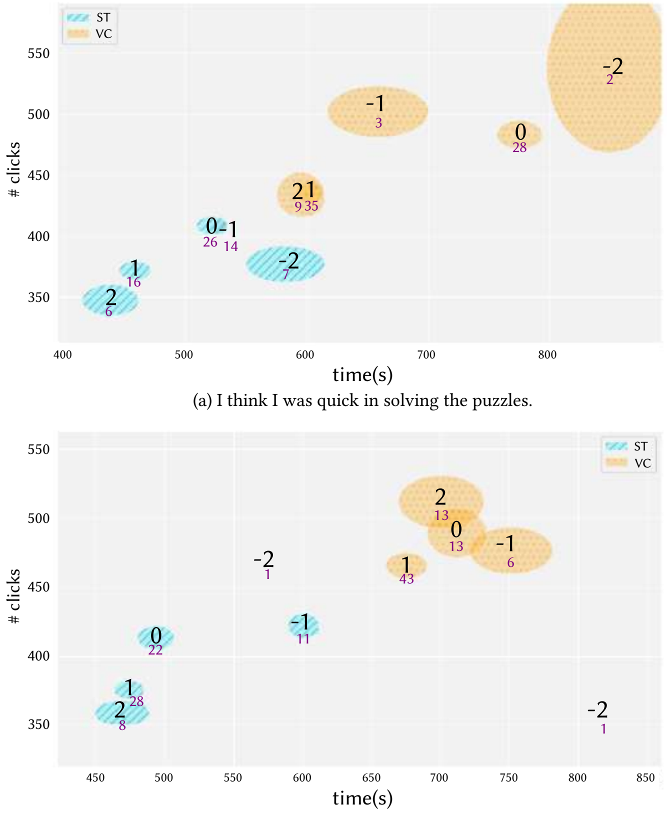


Fig. 11. Which puzzle participants liked the most and the least. Both groups, ST and VC, liked the puzzle C˜ the most, and both liked the puzzle A˜ the least. For some puzzles, there are noticeable differences worth studying (see text for details).



(a) I think I was quick in solving the puzzles.

(b) I found easy to complete the game.

Fig. 12. Means and standard errors (represented as the lengths of the corresponding ellipses' axes) of completion time and number of clicks for VC and ST users, grouped by their Likert replies to whether they (a) think they were quick, and (b) found the game easy to complete. For quicker reading of the plots, the agreement scale is coded as positive (for agreement), negative (for disagreement) and 0 (for neutral), which is the number in the center of the ellipses. The number below this agreement scale is the number of subjects for each particular response.

2. Subjective Perception Vs Actual Effort

Regarding the overall experience (Fig. 10), both groups mostly found it entertaining, compared to boring (Fig. 10a), and surprising, compared to dull (Fig. 10b), but neither exciting nor frustrating (Fig. 10c). However, ST users found it mostly simple, while VC users did not (Fig. 10d). This difference is statistically significant (χ2 = 13.934,

$p = 0.001$) and reinforces the conclusions obtained in the previous sections. Therefore, this can be seen as that users' subjective perception matches objective data. Interestingly, this increase in difficulty does not generally seem to translate into less entertainment or more frustration.

The results of the most and the least preferred puzzles (Fig. 11) are also similar in both groups, despite the actual differences observed regarding actual effort. It is worth noticing how, regardless of the group (ST or VC), puzzles $B^{\sim}$ and $C^{\sim}$ are liked twice and three times as much as their B and C counterparts, respectively, even though they include the additional challenge of distracting images. It is interesting to note that a similar proportion of VC and ST users liked $C^{\sim}$ the most, in spite of $C_{\overset{\sim}{\mathrm{vc}:d}}$ being particularly hard. The puzzle selected as the least liked was A in both ST and VC. When comparing the most and least liked puzzle-wise, it can be observed that some puzzles are judged as the least liked as much ($B_{\overset{\sim}{\mathrm{vc}:c}}$), half as much ($C_{\overset{\sim}{\mathrm{vc}:d}}$), or three times as much ($A_{\overset{\sim}{\mathrm{vc}:e}}$) as they are preferred. These observations lead to two main conclusions: players exhibited a wide variety of skills and tastes, and a notable amount of them seems to have enjoyed some of the most challenging puzzles.

It is important to find out whether users' perception of effort and difficulty align with actual collected metrics. Two questions in the questionnaire (namely, whether they think they were quick, and they found the game easy) relate to the two metrics collected (completion time and number of clicks). Then, it is possible to explore how the answers to these questions distribute in the time-clicks space. Regarding the effort, it can be observed that users' perception strongly correlates with the actual times (Fig. 12a). This happens in both user groups, which is interesting because the time and click ranges are significantly different among these groups. In other words, VC users who felt they were quick, took shorter, on average, than VC users who believed they were not, but longer than ST users in general, even than those ST users who felt they were not quick. It can also be noticed that time correlates better with the user perception of time than the number of clicks.

As for the second question, the perception of difficulty correlates generally better with the number of clicks than with completion times (Fig. 12b). In this case, however, the results are notably different for the VC and ST users. On the one hand, ST users who perceived the puzzles as difficult took generally longer and used more clicks. On the other hand, although the pattern for VC users is possibly less clear, it is intriguing that those VC users who used the most number of clicks, strongly agreed that the game was easy. It can be very tentatively argued that this may partly relate with a state of flow: those users might have been enjoying the game *despite* making more mouse clicks, since for their skills the challenge was at a good difficulty balance. On average, this user subgroup took also longer than those who less strongly claimed to find the game easy.

VI. Discussion

The overall results of this study strongly support the idea that a variety of simple visual ingredients can be used to modulate the game difficulty while preserving the core gameplay. Interestingly, the analysis of user perception through the instant emotional feedback and opinion survey reveals that more challenging puzzles do not necessarily lead to boredom or frustration, but they may also be found entertaining and enjoyable. Although the evidence for claiming that a flow state was achieved by some participants is certainly weak, and not directly pursued at design stage, a door is certainly open regarding this desirable possibility.

Regarding the applicability of this study, puzzles and card games seem the most straightforward choice, which is not little considering that a wide range of these games exist, either as complete games or as mini games. Beyond these genres, however, other possibilities where some variation of this work is feasible includes games whose backgrounds or other visual elements (enemies, non-playable characters) can be modified to modulate the gameplay. It can even be speculated that the concept may inspire not only variations of existing games, but also new games.

The preferences for some puzzles observed in the questionnaire may suggest that solving these puzzles are intrinsically more enjoyable. However, an alternative hypothesis is that these preferences might have to do with the particular images being used. Further work might explore the use of the same images across different visual concepts to remove this potential bias, and making sure that the questions are understood as meant. Regarding the different visual computing elements considered, they produce a variety of effects and it is therefore hard to generalize how exactly each contributes to modulate the difficulty. All in all, however, the combination of abstract shapes and the dynamic effect (rotational motion) led to more participants to surrender, and therefore this visual effect should be used only when the highest challenge is advisable. The color-based modification has possibly contributed to lower the difficulty to the extent that using distracting images neutralized its effect. This suggests that the same visual concept may have a notably different impact depending on which images it is applied to. The edge-like effect was possibly the least popular and among the hardest ones, both with similar and dissimilar distracting images. Overall, the challenge that a particular visual element will induce has to be carefully studied, and it is possible hard to predict. One possibility to address this is relying on the body of knowledge of the human cognition and visual perception, to leverage on its strengths and weaknesses. Another possibility would be to automatically learn the implications of each of a large set of visual effects through large-scale long-term usage.

User feedback is provided when a row or column of cubes are correctly completed. Indirectly, this feedback acts as a facilitating mechanism and, while this is applied to all puzzles, it makes the hardest puzzles easier, which complicates the derivation of accurate conclusions on how hard the visual concept itself is, since the feedback may be interfering. In addition, this feedback might also promote some players to apply some form of strategic playing. Therefore, these issues would require further examination and experimentation.

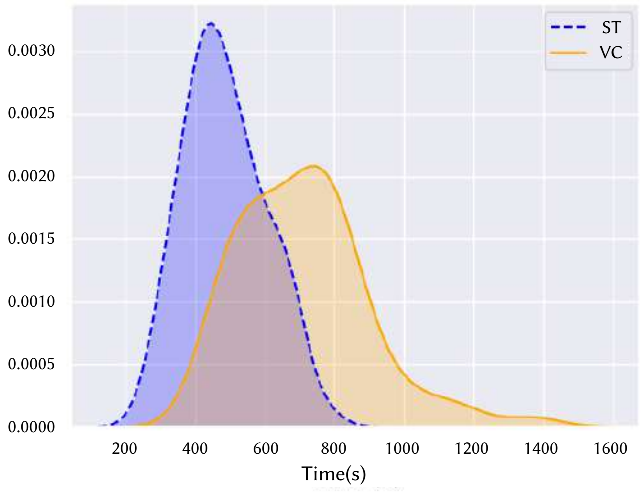
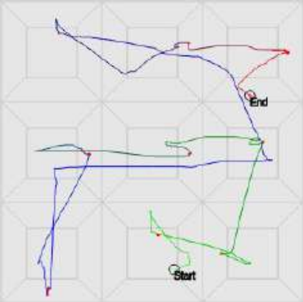
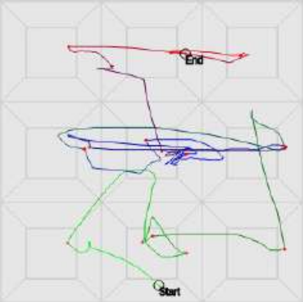
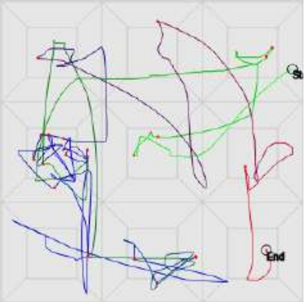
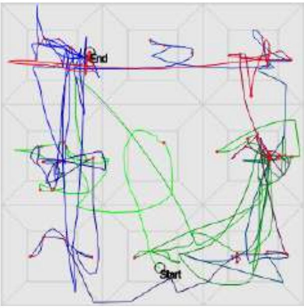
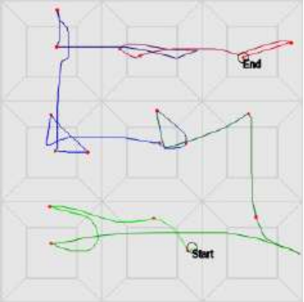
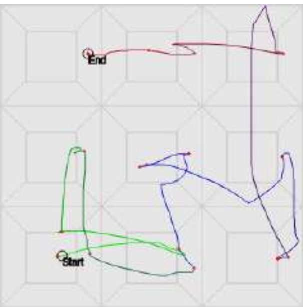
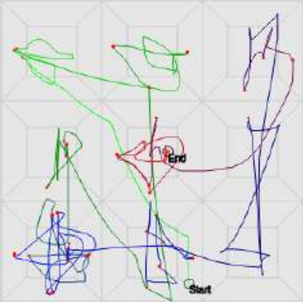
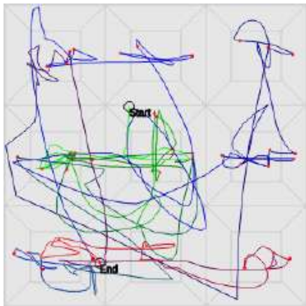
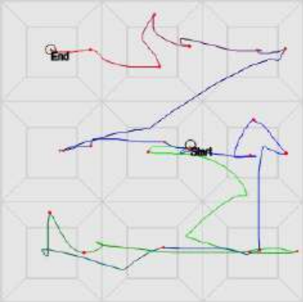
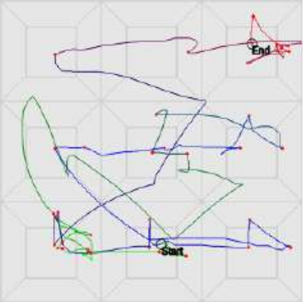
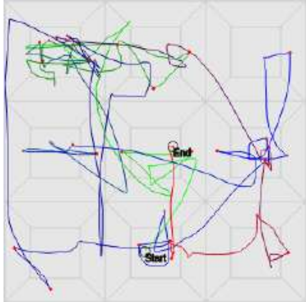
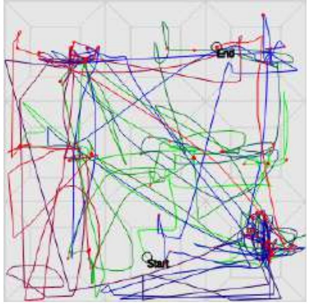


Fig. 13. Density estimation of time for game completions in VC and ST. The distributions for the number of clicks are similar.

TABLE XIV. Mouse Traces and Clicks for Each of the Six Puzzles as Solved by Two Participants, One Under ST and Another Under VC Who, Respectively, Took the Least and the Most to Complete the Game. For Each Puzzle, the Start and End Cursor Points are Labeled, and the Percent of Elapsed Time is Represented With the Trace Color (From Green to Blue to Purple to Red). The Nine Pieces (Cubes) and the Five Areas Per Cube Used to Rotate it are Illustrated in the Background as a Visualization Guide. This Representation Provides an Idea of the Effort Devoted to a Particular Puzzle by a Particular User, and it Lends Itself to Per-Puzzle, Per-User, and Per-Condition Analyses

	User under ST		User under VC	
	distinct	alike	distinct	alike
A				
B				
C				

It is interesting to note that the number of clicks and the elapsed time to complete individual puzzles correlate well with the particular game version that users played. This implies that these metrics may have some potential predictive power of how hard players are finding a game or part of it. However, the time distributions (Fig. 13) reveal two important observations. First, despite having statistically different means, there is a notable distribution overlap; second, the time distribution for VC players is right long-tailed (in fact, this distribution is not normal whereas that of ST is). These observations mean that many VC users actually took shorter than the average VC user, with times close to ST users, and that some VC others took significantly longer. Therefore, although VC is *generally* and intrinsically harder, user skill is also a critical factor in explaining the observed performance.

The considerations above suggest that additional data, besides times and clicks, can be useful for player modeling and predicting their skills. For instance, mouse traces (Table XIV) could bring an additional metric of effort (e.g. length of mouse trajectories) and, more generally, they can be very insightful into the different solution strategies employed by users, or for guessing which sense of control over the game a particular user is having.

Beyond manual inspection of these visualizations, or deriving heuristic metrics, disciplined machine learning techniques can turn out to be very useful for characterizing players and, in turn, dynamic game adaptation. In fact, a first approach to this problem has recently been explored [85].

VII. Conclusions

This work has empirically shown that simple visual computing techniques can be introduced to modulate the play experience *without* any other change in the gameplay. In particular, the conducted between-subjects user study revealed significant performance differences between users who played the visually-modified game and those who played a standard, unmodified game version. Interestingly, the user feedback in terms of in-game emojis and a final questionnaire, suggest that the more challenging parts of the games are not necessarily always found frustrating, but can also be found enjoyable by some users.

The work has used three visual concepts (edge, color mapping, and rotating motion), plus distracting images looking similar to the target image, on a cube puzzle web game prototype. Further work may explore other visual computations and alternative games to keep exploring this research theme. The findings of this first study can inform the design of games where these kind of visual modifications can be easily introduced, as well as guide further related research. We believe that through data obtained from players' interaction, machine learning techniques can be leveraged to model users and, eventually, be able to adapt dynamically the game to match each individual's skills. In the long term, this would lead to more enjoyable game experiences.

Acknowledgment

We are very grateful to all participants in the study. The contribution of Francisco Andújar De Andrés in the implementation of the mouse traces is acknowledged.

References

[1] J. H. Brockmyer, C. M. Fox, K. A. Curtiss, E. McBroom, K. M. Burkhart, J. N. Pidruzny, "The development of the game engagement questionnaire: A measure of engagement in video game-playing," *Journal of Experimental Social Psychology*, vol. 45, no. 4, pp. 624–634, 2009.

[2] E. Knekta, C. Runyon, S. Eddy, "One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research," *CBE life sciences education*, vol. 18, no. 1, 2019.

[3] D. Johnson, M. J. Gardner, R. Perry, "Validation of two game experience scales: The player experience of need satisfaction (PENS) and game experience questionnaire (GEQ)," *International Journal of Human-Computer Studies*, vol. 118, pp. 38–46, 2018.

[4] A. Denisova, C. Guckelsberger, D. Zendle, "Challenge in digital games: Towards developing a measurement tool," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, 2017, pp. 2511–2519.

[5] A. Denisova, P. Cairns, C. Guckelsberger, D. Zendle, "Measuring perceived challenge in digital games: Development & validation of the challenge originating from recent gameplay interaction scale (CORGIS)," *International Journal of Human-Computer Studies*, vol. 137, 2020.

[6] M. Csikszentmihalyi, *Flow: The Psychology of Optimal Experience.* Harper Perennial, 1991.

[7] B. Cowley, D. Charles, M. Black, R. Hickey, "Toward an understanding of flow in video games," *Computers in Entertainment*, vol. 6, July 2008.

[8] K. Jalife, C. Holmgård, "Cognitive components of flow states and applications to video game design: A brief framework," in *Proceedings of the 2019 ACM SIGCHI Conference on Creativity and Cognition, C&C 2019, San Diego, CA, USA, June 23-26, 2019*, 2019, pp. 570–577, ACM.

[9] P. Sweetser, P. Wyeth, "Gameflow: A model for evaluating player enjoyment in games," *Computers in Entertainment*, vol. 3, July 2005.

[10] K. Procci, A. R. Singer, K. R. Levy, C. Bowers, "Measuring the flow experience of gamers: An evaluation of the DFS-2," *Computers in Human Behavior*, vol. 28, no. 6, pp. 2306– 2312, 2012.

[11] C. Shu-Hui, W. Wann-Yih, J. Dennison, "Validation of EGameFlow: A self-report scale for measuring user experience in video game play," *Computers in Entertainment*, vol. 16, Sept. 2018.

[12] P. Sweetser, D. Johnson, P. Wyeth, A. Anwar, Y. Meng, A. Ozdowska, "Gameflow in different game genres and platforms," *Computers in Entertainment*, vol. 15, Apr. 2017.

[13] M.-Y. Hwang, J.-C. Hong, Y. wei Hao, J.-T. Jong, "Elders' usability, dependability, and flow experiences on embodied interactive video games," *Educational Gerontology*, vol. 37, no. 8, pp. 715–731, 2011.

[14] M.-J. Tsai, L.-J. Huang, H.-T. Hou, C.-Y. Hsu, G.-L. Chiou, "Visual behavior, flow and achievement in game-based learning," *Computers & Education*, vol. 98, pp. 115–129, 2016.

[15] Y.-H. Hsieh, Y.-C. Lin, H.-T. Hou, "Exploring the role of flow experience, learning performance and potential behavior clusters in elementary students' game-based learning," *Interactive Learning Environments*, vol. 24, no. 1, pp. 178– 193, 2016.

[16] Y.-C. Yeh, S.-Y. Chen, E. M. Rega, C.-S. Lin, "Mindful learning experience facilitates mastery experience through heightened flow and self-efficacy in game-based creativity learning," *Frontiers in Psychology*, vol. 10, 2019.

[17] M. T. F. Mak, M. Wang, K. W. S. Chu, "Effects of a gamified learning platform on elementary school students' flow experiences in leisure reading," *Proceedings of the Association for Information Science and Technology*, vol. 56, no. 1, pp. 454–458, 2019.

[18] M. Zheng, H. A. Spires, *Fifth Graders' Flow Experience in a Digital Game-Based Science Learning Environment*, pp. 1433–1450. I. Management Association, 2015.

[19] X. Fang, S. Chan, J. Brzezinski, C. Nair, "Development of an instrument to measure enjoyment of computer game play," *International Journal of Human–Computer Interaction*, vol. 26, no. 9, pp. 868–886, 2010.

[20] A. Corcos, "Being enjoyably challenged is the key to an enjoyable gaming experience: an experimental approach in a first-person shooter game," *Socioaffective Neuroscience & Psychology*, vol. 8, no. 1, 2018.

[21] A. Touati, Y. Baek, "What leads to player's enjoyment and achievement in a mobile learning game?," *Journal of Educational Computing Research*, vol. 56, no. 3, pp. 344– 368, 2018.

[22] Y. Baek, A. Touati, "Exploring how individual traits influence enjoyment in a mobile learning game," *Computers in Human Behavior*, vol. 69, pp. 347–357, 2017.

[23] C. Jennett, A. L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs, A. Walton, "Measuring and defining the experience of immersion in games," *International Journal of Human-Computer Studies*, vol. 66, no. 9, pp. 641–661, 2008.

[24] J. Chen, "Flow in games (and everything else)," *Communications of the ACM*, vol. 50, pp. 31–34, Apr. 2007.

[25] R. C. Gonzalez, R. E. Woods, *Digital Image Processing*. Prentice Hall, 2018. 4th. edition.

[26] W. Thompson, R. Fleming, S. Creem-Regehr, J. Kelly Stefanucci, *Visual Perception from a Computer Graphics Perspective.* A K Peters/CRC Press, 2011.

[27] A. M. Barry, *Visual intelligence: perception, image, and manipulation in visual communication.* Albany: State University of New York Press, 1997.

[28] M. S. El-Nasr, S. Yan, "Visual attention in 3D video games," in *Proceedings of the ACM SIGCHI International Conference on Advances in Computer Entertainment Technology*, 2006, p. 22–es, ACM.

[29] R. Dubey, P. Agrawal, D. Pathak, T. Griffiths, A. A. Efros, "Investigating human priors for playing video games," in *International Conference on Machine Learning (ICML)*, July 2018, pp. 1348–1356.

[30] C. Green, D. Bavelier, "Action-video-game experience alters the spatial resolution of vision," *Psychological Science*, vol. 18, no. 1, pp. 88–94, 2007.

[31] Y.-H. Kim, D.-W. Kang, D. Kim, H.-J. Kim, Y. Sasaki, T. Watanabe, "Real-time strategy video game experience and visual perceptual learning," *Journal of Neuroscience*, vol. 35, pp. 10485–10492, July 2015.

[32] W. Boot, D. Blakely, D. Simons, "Do action video games improve perception and cognition?," *Frontiers in Psychology*, vol. 2, no. SEP, 2011.

[33] R. L. Achtman, C. S. Green, D. Bavelier, "Video games as a tool to train visual skills.," *Restorative neurology and neuroscience*, vol. 26 4-5, pp. 435–446, 2008.

[34] I. Spence, J. Feng, "Video games and spatial cognition," *Review of General Psychology*, vol. 14, no. 2, pp. 92–104, 2010.

[35] D. Kersten, "Perception, computer graphics, and video games," *Journal of Vision*, vol. 7, no. 15, 2007.

[36] D. Dziedzic, W. Włodarczyk, "Approaches to measuring the difficulty of games in dynamic difficulty adjustment systems," *International Journal of Human–Computer Interaction*, vol. 34, no. 8, pp. 707–715, 2018.

[37] O. Sotamaa, "When the game is not enough: Motivations and practices among computer game modding culture," *Games and Culture*, vol. 5, no. 3, pp. 239–255, 2010.

[38] J. Schell, *The Art of Game Design: A Book of Lenses.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008.

[39] "Mirror's Edge." https://en.wikipedia.org/wiki/ Mirror's_Edge.

[40] J. Juul, *A Casual Revolution: Reinventing Video Games and Their Players.* The MIT Press, 2009.

[41] K. Hicks, P. Dickinson, J. Holopainen, K. Gerling, "Good game feel: An empirically grounded framework for juicy design," in *Proceedings of the 2018 DiGRA International Conference: The Game is the Message*, July 2018, DiGRA.

[42] E. Buckthal, "Juiciness in citizen science computer games: Analysis of a prototypical game," Master's thesis, 2014.

[43] J. Juul, B. Jason, "Good feedback for bad players? A preliminary Study

of ‘juicy’ Interface feedback,” in *First joint FDG/DiGRA Conference*, 2016.

[44] J. S. Prook, D. P. Janssen, S. Gualeni, “The negative effects of praise and flattery,” in *Foundations of Digital Games*, 2015.

[45] J. Korinek, S. L. Fagerli, “Juiciness — a study of visual effects in games,” Master’s thesis, Aalborg University, 2017.

[46] K. Hicks, K. Gerling, G. Richardson, T. Pike, O. Burman, P. Dickinson, “Understanding the effects of gamification and juiciness on players,” in *IEEE Conference on Games (CoG)*, 2019.

[47] K. Hicks, K. Gerling, P. Dickinson, V. Vanden Abeele, “Juicy game design: Understanding the impact of visual embellishments on player experience,” in *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY ’19, 2019, pp. 185–197.

[48] D. Kao, “The effects of juiciness in an action RPG,” *Entertainment Computing*, vol. 34, 2020.

[49] F. James, “Crash 4 n. verted mode explained: How it works and what to expect with the inverted version of levels,” Oct. 2020. https://www.gamesradar.com/crash-4-nverted-mode.

[50] “Mirror mode (Mario Kart Racing Wiki).” https:// mariokart.fandom.com/wiki/Mirror_Mode.

[51] D. Kao, D. F. Harrell, “Toward understanding the impact of visual themes and embellishment on performance engagement, and self-efficacy in educational games,” in *American Educational Research Association Conference*, 2017.

[52] D. Kao, D. F. Harrell, “Exploring the impact of avatar color on game experience in educational games,” in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA ’16, 2016, pp. 1896– 1905.

[53] “Colorblind mode (Battlefield wiki).” https:// battlefield.fandom.com/wiki/Colorblind_Mode.

[54] A. Rodríguez-Benítez, I. Boada, S. Thió-Henestrosa, M. Sbert, “Cprforblind: A video game to introduce cardiopulmonary resuscitation protocol to blind people,” *British Journal of Educational Technologies*, vol. 49, no. 4, pp. 636–645, 2018.

[55] M. Magdics, C. Sauvaget, R. J. García, M. Sbert, “Post-processing NPR effects for video games,” in *Proceedings of the 12th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*, VRCAI ’13, 2013, pp. 147–156, Association for Computing Machinery.

[56] “Non-photorealistic rendering.” https://en.wikipedia.org/wiki/Non-photorealistic_rendering.

[57] D. Gravina, A. Khalifa, A. Liapis, J. Togelius, G. N. Yannakakis, “Procedural content generation through quality diversity,” in *IEEE Conference on Games (CoG)*, 2019, pp. 1– 8.

[58] M. Zohaib, “Dynamic Difficulty Adjustment (DDA) in computer games: A review,” *Advances in Human-Computer Interaction*, vol. 2018, 2018.

[59] T. Constant, G. Levieux, “Dynamic difficulty adjustment impact on players’ confidence,” in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2019, ACM.

[60] A. Denisova, P. Cairns, “Player experience and deceptive expectations of difficulty adaptation in digital games,” *Entertainment Computing*, vol. 29, pp. 56–68, 2019.

[61] T. A. Sanocki, K. W. Bowyer, M. Heath, S. Sarkar, “Are edges sufficient for object recognition?,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 24, pp. 340–349, Jan. 1998.

[62] M. A. Webster, K. Halen, A. J. Meyers, P. Winkler, J. S. Werner, “Colour appearance and compensation in the near periphery,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 277, Feb. 2010.

[63] D. K. M. G. Bloj, A. C. Hurlbert, “Perception of three-dimensional shape influences colour perception through mutual illumination,” *Nature*, vol. 402, pp. 877–879, 1999.

[64] L. Teichmann, T. Grootswagers, T. A. Carlson, A. N. Rich, “Seeing versus knowing: The temporal dynamics of real and implied colour processing in the human brain,” *NeuroImage*, vol. 200, pp. 373–381, 2019.

[65] J. Pan, G. Bingham, “With an eye to low vision: Optic flow enables perception despite image blur,” *Optometry and Vision Science*, vol. 90, pp. 1119–1127, Oct. 2013.

[66] P. Lewis, R. Rosén, P. Unsbo, J. Gustafsson, “Resolution of static and dynamic stimuli in the peripheral visual field,” *Vision Research*, vol. 51, no. 16, pp. 1829–1834, 2011.

[67] J. Maarseveen, C. L. Paffen, F. A. Verstraten, H. Hogendoorn, “Representing dynamic stimulus information during occlusion,” *Vision Research*, vol. 138, pp. 40–49, 2017.

[68] F. Ghayour, D. Cantor, *Real-Time 3D Graphics with WebGL 2: Build Interactive 3D Applications with JavaScript and WebGL 2 (OpenGL ES 3.0), 2nd Edition.* Packt Publishing, 2018.

[69] S. K. Bhatia, V. Lakshminarayanan, A. Samal, G. V. Welland, “Human face perception in degraded images,” *Journal of Visual Communication and Image Representation*, vol. 6, no. 3, pp. 280–295, 1995.

[70] M. A. Webster, D. I. A. MacLeod, “Visual adaptation and face perception,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 366, June 2011.

[71] D. P. Katharina Dobs, Leyla Isik, N. Kanwisher, “How face perception unfolds over time,” *Nature Communications*, vol. 10, 2019.

[72] D. Patterson, “The human face in play based, shared, digital learning experiences,” in *Proceedings of the Australasian Computer Science Week Multiconference*, 2019, ACM.

[73] T. Bachmann, *Perception of Pixelated Images.* Academic Press, 2016.

[74] N. Bowman, T. Wulf, “Finding nostalgia in the pixelated video games of decades past,” Aug. 2018. https:// theconversation.com/finding-nostalgia-in-the-pixelated-video-games-of-decades-past-98518.

[75] T. Wulf, N. D. Bowman, J. A. Velez, J. Breuer, “Once upon a game: Exploring video game nostalgia and its impact on well-being,” *Psychology of Popular Media Culture*, vol. 9, no. 1, pp. 83–95, 2018.

[76] C.-C. Carbon, “Understanding human perception by human-made illusions,” *Frontiers in Human Neuroscience*, vol. 8, July 2014.

[77] R. M. Williams, R. V. Yampolskiy, “Optical illusions images dataset,” *CoRR*, vol. abs/1810.00415, 2018.

[78] L. E. Nacke, “An introduction to physiological player metrics for evaluating games,” in *Game Analytics, Maximizing the Value of Player Data*, 2013, pp. 585–619.

[79] G. Chanel, C. Rebetez, M. Bétrancourt, T. Pun, “Emotion assessment from physiological signals for adaptation of game difficulty,” *IEEE Transactions on Systems, Man, and Cybernetics -Part A: Systems and Humans*, vol. 41, no. 6, pp. 1052–1063, 2011.

[80] P. Dragicevic, “Fair Statistical Communication in HCI,” in *Modern Statistical Methods for HCI*, Springer, 2016, pp. 291– 330.

[81] J. B. Vornhagen, A. Tyack, E. D. Mekler, “Statistical significance testing at CHI PLAY: Challenges and opportunities for more transparency,” in *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY ’20, New York, NY, USA, 2020, pp. 4–18, Association for Computing Machinery.

[82] P. Dragicevic, *Fair Statistical Communication in HCI*, pp. 291–330. Springer International Publishing, 2016.

[83] P. D. Ellis, *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results.* Cambridge University Press, 2010.

[84] K. Jung, J. Lee, V. Gupta, G. Cho, “Comparison of bootstrap confidence interval methods for GSCA using a Monte Carlo simulation,” *Frontiers in Psychology*, vol. 10, 2019.

[85] X. Anadón, P. Sanahuja, V. J. Traver, A. Lopez, J. Ribelles, “Characterising players of a cube puzzle game with a two-level bag of words,” in *Adjunct Publication of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, Utrecht, The Netherlands, June 21-25, 2021*, 2021, pp. 47–53, ACM.

Jose Ribelles

He is Associate Professor of the Computer Languages & Systems department and member of the Institute of New Imaging Technologies, both at the Universitat Jaume I of Castellon, Spain. He earned a PhD in Computer Science in 2000. His teaching covers subjets as Game Character Modeling, Computer Animation, or Image Synthesis. His research has focused on Computational Photography, Computer Graphics, and, more rencently, Videogames and Human-Computer Interaction.

Angeles Lopez

She has a PhD on Computer Science (Universitat Jaume I de Castellón, 1999) and is an Associate Professor in the Departament of Engineering and Science of Computers since 2002. She is a member of the Visual Engineering research group, integrated in the Institute of New Imaging Technologies (INIT). Her current research interests include depth estimation from stereo pair, single image or spherical panorama, and their applications in Computer Vision, Virtual Reality and Computational Photography.

V. Javier Traver

He is a member of the Computer Languages & Systems department at Universitat Jaume I, where he lectures at several undergraduate degrees (Computer Science, Videogame Design and Development, Intelligent Robotics) and at master level (Intelligent Systems). He is also affiliated with the Institute of New imaging Technologies (INIT), and his research interests include a broad range of topics within Computer Vision, Machine Learning, Human-Computer Interaction, and Videogames.

unir
LA UNIVERSIDAD
EN INTERNET

Rectorado
Avenida de la Paz, 137
26006 Logroño (La Rioja)
t (+34) 941 21 02 11

www.unir.net