

International Journal of
Interactive Multimedia
and Artificial Intelligence

December 2023, Vol. VIII, Number 4
ISSN: 1989-1660

unir LA UNIVERSIDAD
EN INTERNET

*“The program IS the rules, and regardless of
where they came from, it is those rules – that is,
the program – that generates material I could
never have imagined or generated myself.”*

Harold Cohen

EDITORIAL TEAM

Editor-in-Chief

Dr. Elena Verdú, Universidad Internacional de La Rioja (UNIR), Spain

Dr. Rubén González Crespo, Universidad Internacional de La Rioja (UNIR), Spain

Managing Editors

Dr. Xiomara Patricia Blanco Valencia, Universidad Internacional de La Rioja (UNIR), Spain

Dr. Paulo Alonso Gaona-García, Universidad Distrital Francisco José de Caldas, Colombia

Office of Publications

Editorial Coordination

Dr. Pedro Hípola, Universidad Internacional de La Rioja (UNIR), Spain

Lic. Blanca Albarracín, Universidad Internacional de La Rioja (UNIR), Spain

Indexing and Metrics

Dr. Álvaro Cabezas Clavijo, Universidad Internacional de La Rioja (UNIR), Spain

Lic. Mercedes Contreras, Universidad Internacional de La Rioja (UNIR), Spain

Layout and graphic edition

Lic. Ainhoa Puente, Universidad Internacional de La Rioja (UNIR), Spain

Associate Editors

Dr. Enrique Herrera-Viedma, University of Granada, Spain

Dr. Witold Perdrycz, University of Alberta, Canada

Dr. Kuan-Ching Li, Providence University, Taiwan

Dr. Robertas Damaševičius, Kaunas University of Technology, Lithuania

Dr. Javier Martínez Torres, Universidad de Vigo, Spain

Dr. Miroslav Hudec, University of Economics of Bratislava, Slovakia

Dr. Seifedine Kadry, Noroff University College, Norway

Dr. Nilanjan Dey, Techno International New Town, India

Dr. Mahdi Khosravy, Cross Labs, Cross Compass Ltd., Tokyo, Japan

Dr. Jörg Thomaschewski, Hochschule Emden/Leer, Emden, Germany

Dr. Mu-Yen Chen, National Cheng Kung University, Taiwan

Dr. Francisco Mochón Morcillo, National Distance Education University, Spain

Dr. Manju Khari, Jawaharlal Nehru University, New Delhi, India

Dr. Carlos Enrique Montenegro Marín, Francisco José de Caldas District University, Colombia

Dr. Yaping Mao, Qinghai Normal University, China

Dr. Juan Manuel Corchado, University of Salamanca, Spain

Dr. Giuseppe Fenza, University of Salerno, Italy

Dr. S.P. Raja, Vellore Institute of Technology, Vellore, India

Dr. Jerry Chun-Wei Lin, Western Norway University of Applied Sciences, Norway

Dr. Juan Antonio Morente, University of Granada, Spain

Dr. Abbas Mardani, The University of South Florida, USA

Dr. Amrit Mukherjee, University of South Bohemia, Czech Republic

Dr. José Ignacio Rodríguez Molano, Universidad Distrital Francisco José de Caldas, Colombia

Dr. Suyel Namasudra, National Institute of Technology Agartala, India

Editorial Board Members

Dr. Rory McGreal, Athabasca University, Canada

Dr. Óscar Sanjuán Martínez, Lumen Technologies, USA

Dr. Anis Yazidi, Oslo Metropolitan University, Norway

Dr. Juan Pavón Mestras, Complutense University of Madrid, Spain

Dr. Smriti Srivastava, Netaji Subhas University of Technology, New Delhi, India

Dr. Lei Shu, Nanjing Agricultural University, China/University of Lincoln, UK

Dr. Ali Selamat, Malaysia Japan International Institute of Technology, Malaysia
Dr. Hamido Fujita, Iwate Prefectural University, Japan
Dr. Francisco García Peñalvo, University of Salamanca, Spain
Dr. Francisco Chiclana, De Montfort University, United Kingdom
Dr. S. Vimal, Ramco Institute of Technology, Tamil Nadu, India
Dr. Jordán Pascual Espada, Oviedo University, Spain
Dr. Ioannis Konstantinos Argyros, Cameron University, USA
Dr. Ligang Zhou, Macau University of Science and Technology, Macau, China
Dr. Palanichamy Naveen, KPR Institute of Engineering and Technology, India
Dr. Juan Manuel Cueva Lovelle, University of Oviedo, Spain
Dr. Pekka Siirtola, University of Oulu, Finland
Dr. Peter A. Henning, Karlsruhe University of Applied Sciences, Germany
Dr. Yago Saez, Universidad Carlos III de Madrid, Spain
Dr. Vijay Bhaskar Semwal, National Institute of Technology, Bhopal, India
Dr. Anand Paul, Kyungpook National University, South Korea
Dr. Javier Bajo Pérez, Polytechnic University of Madrid, Spain
Dr. Jinlei Jiang, Dept. of Computer Science & Technology, Tsinghua University, China
Dr. B. Cristina Pelayo G. Bustelo, University of Oviedo, Spain
Dr. Masao Mori, Tokyo Institute of Technology, Japan
Dr. Rafael Bello, Universidad Central Marta Abreu de Las Villas, Cuba
Dr. Daniel Burgos, Universidad Internacional de La Rioja - UNIR, Spain
Dr. JianQiang Li, Beijing University of Technology, China
Dr. Rebecca Steinert, RISE Research Institutes of Sweden, Sweden
Dr. Monique Janneck, Lübeck University of Applied Sciences, Germany
Dr. Carina González, La Laguna University, Spain
Dr. Mohammad S Khan, East Tennessee State University, USA
Dr. David L. La Red Martínez, National University of North East, Argentina
Dr. Juan Francisco de Paz Santana, University of Salamanca, Spain
Dr. José Estrada Jiménez, Escuela Politécnica Nacional, Ecuador
Dr. Octavio Loyola-González, Stratesys, Spain
Dr. Guillermo E. Calderón Ruiz, Universidad Católica de Santa María, Peru
Dr. Moamin A Mahmoud, Universiti Tenaga Nasional, Malaysia
Dr. Madalena Riberio, Polytechnic Institute of Castelo Branco, Portugal
Dr. Manik Sharma, DAV University Jalandhar, India
Dr. Edward Rolando Núñez Valdez, University of Oviedo, Spain
Dr. Juha Röning, University of Oulu, Finland
Dr. Paulo Novais, University of Minho, Portugal
Dr. Sergio Ríos Aguilar, Technical University of Madrid, Spain
Dr. Hongyang Chen, Fujitsu Laboratories Limited, Japan
Dr. Fernando López, Universidad Complutense de Madrid, Spain
Dr. Runmin Cong, Beijing Jiaotong University, China
Dr. Manuel Perez Cota, Universidad de Vigo, Spain
Dr. Abel Gomes, University of Beira Interior, Portugal
Dr. Víctor Padilla, Universidad Internacional de La Rioja - UNIR, Spain
Dr. Mohammad Javad Ebadi, Chabahar Maritime University, Iran
Dr. Andreas Hinderks, University of Sevilla, Spain
Dr. Brij B. Gupta, National Institute of Technology Kurukshetra, India
Dr. Alejandro Baldominos, Universidad Carlos III de Madrid, Spain

Editor's Note

THE International Journal of Interactive Multimedia and Artificial Intelligence – IJIMAI – 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. The present regular issue comprises different topics as generative AI, brain and main inspired computing, bird species identification, spam detection, recommendation systems, synthetic aperture radar automatic target recognition, hand gestures recognition, anomalies detection for video surveillance systems, disease detection, social networks analysis, or user experience. The collection of articles shows the wide use of deep learning techniques, although classical machine learning techniques, among others, are also present.

The issue starts with a topic that is getting a lot of attention from various media due to its popularity, because we can all easily use tools like chatGPT, Midjourney or Dall-E. These tools are considered generative AI that can generate content very similar to that generated by humans. The first article by Garcia-Peñalvo and Vázquez-Ingelmo, presents a literature mapping of AI-driven content generation, analyzing 631 solutions published over the last five years to better understand and characterize the generative AI landscape. Due to the concerns and acceptance issues that have arisen in society as a result of the emergence of this technology, the authors suggest more comprehensive understanding of what generative AI entails, so that the potential challenges are addressed more pragmatically and effectively.

The second article of this issue focuses on brain and mind inspired computing (BMC), an emerging research field. Liu and Wei design a model and framework for BMI theory. Based on the brain mechanism and mind architecture, they propose a semantic-oriented multimedia neural, cognitive computing model for multimedia semantic computing. Besides, they propose a hierarchical cross-modal cognitive neural computing framework for cross-modal information processing. Moreover, they propose a cross-modal neural, cognitive computing architecture for a remote sensing intelligent information extraction platform and unmanned autonomous system. Their research on remote sensing intelligent information extraction and cross-media information retrieval shows that the scene classification, target detection, target classification and target recognition based on the BMC algorithm work. The BMI theory proposed can be widely used in high-resolution earth observation systems and many other applications.

Next article describes a solution based on convolutional neural networks for bird species identification, which is crucial for avian diversity conservation. Das et al. explore the ability of deep transfer models such as VGG16, VGG19 and InceptionV3 to effectively extract and discriminate speech signals from different species of birds. The networks were trained using data from 37 bird species, obtaining accuracies equal to 78, 61.9 and 85%, respectively. In practical terms, the suggested method may be of great use to ornithologists by making the identification of bird species a straightforward process.

Also using convolutional neural networks, Vélez de Mendizabal et al. propose a new technique to decode Leetspeak. This is a type of slang writing that replaces some characters with visually similar symbols, preventing the spam classifiers recognising words, so that the spam emails are not detected. When messages are deobfuscated, the performance of the classifiers increases and reaches, in many cases, the values obtained when messages have not been obfuscated.

The authors propose a reliable convolutional neural network (CNN) design for Leetspeak deobfuscation processes and its evaluation, an image database used for training, and four datasets for evaluating Leetspeak deobfuscation processes.

Another type of neural networks, gated graph neural networks, are used in the next article. Seo and Kim target the problem of sequential recommendation to predict user's next action based on personal action sequences. Data sparsity is a challenge in these problems and translation-based recommendations, which learn distance metrics to capture interactions between users and items in sequential recommendations, contributes to overcome this issue. The authors propose an attentive flexible translation for recommendations (AFTRec) to predict the user's next item in sparse sequential recommendation datasets. Experiments using four sparse datasets and one dense dataset with different domains show that AFTRec outperforms the state-of-the-art baselines in terms of normalized discounted cumulative gain and hit rate on sparse datasets.

By also using attention mechanisms, Ukwuoma et al. target the problem of synthetic aperture radar (SAR) automatic target recognition (ATR). Their paper introduces a new attention based ResNet architecture appropriate for the SAR recognition task. They propose a simple channel attention mechanism into a ResNet architecture for SAR ATR involving only a handful of parameters while attaining clear performance gains. They also explore the One Policy Learning Rate on the ResNet-50 architecture and compare it with the proposed attention based ResNet-50 architecture. With the attention based model and the One Policy Learning Rate-based architecture, they were able to obtain recognition rates of 100% and 99.8%, respectively.

Next paper presents a solution that allows the recognition of hand gestures by analyzing three-dimensional landmarks using also deep learning technology. Osimani et al. propose the identification of 9 hand gestures by interpreting a cloud of 3D reference points obtained through a standard RGB camera. They introduce a neural network architecture that has a small number of hidden layers but high prediction hit rate of hand gestures. One of the main contributions, that considerably improves the performance, is a first layer of normalization and transformation of the landmarks. In their experimental analysis, they reach an accuracy of 99.87% recognizing 9 hand gestures.

In the field of video surveillance, Qasim Gandapur and Verdú present an automated deep learning model that detects and prevents anomaly activities. A real-world video surveillance system is designed based on the ResNet-50 architecture to extract the high-level features from input streams, while temporal features are extracted by a convolutional gated recurrent unit from the ResNet-50 extracted features in the time-series dataset. The UCF-Crime dataset is used to evaluate the proposed deep learning model, achieving 82.22% accuracy. In addition, the proposed model outperforms related deep learning models.

Next article, by Mariammal et al., examines the performance of classical machine learning methods as bagging, random forest, support vector machine, decision tree, Naïve Bayes and k-nearest neighbor classifiers using a crop dataset, a prisoners' dataset and the iris dataset. The results show that the bagging ensemble technique outperforms the rest.

Torres et al. focus on the development of a system for the detection of downy mildew disease in roses through image analysis using convolutional neural networks and the correlation of environmental

variables through an experiment in a controlled environment. Besides, they develop an IoT platform that integrates the artificial intelligence module. The model is validated comparing with three different models of neural networks. According to the tests and the analysis of the results obtained with the microclimatic variables, it is observed that the relative humidity variable can influence the development and appearance of downy mildew disease when its value is above 85% during an extended period of the system.

Also targeting disease detection, the article by Intriago-Pazmiño et al. describes a study of four metrics to find their best parameters. The metrics are Contrast Improvement Index (CII), Enhancement Measurement Estimation (EME), Entropy EME (EMEE), and Entropy. The metrics are applied to two cases of studies: fundus and mammography images, on five datasets. These datasets contain healthy and pathological images, and some are poor quality images. Based on the experimental results provided, the conclusion is that EMEE, EME, and CII metrics are valuable for measuring the enhancement of the studied medical images.

A clustering validation index (CVI) is a metric used to evaluate the results of a clustering algorithm. In the next article, Kumar Sharma et al. propose an internal CVI to be used as a complementary measure to the available internal CVIs. These are used frequently in clustering to measure the goodness of the clustering algorithms without taking any external inputs. The proposed index uses a modified compactness measure and an updated separation measure, based on the notion of S-divergence. A total of 10 databases of two classes, synthetic and realworld ones, are considered in this work to prove the effectiveness of the proposed metric over some of the most popular existing internal CVIs. Empirical results with four popular clustering algorithms show that the proposed index is proficient in determining the number of clusters and the best partition for several of the databases, including the database with arbitrary cluster shapes.

In the following paper, Debbi proposes a causal explanation technique for conjunctive queries in probabilistic databases. While query answer explanation in relational databases focuses on why is a tuple in a query result, in probabilistic databases, it should also explain why the tuple has a certain probability. Based on the notions of causality, responsibility and blame, the author addresses explanation for tuple and attribute uncertainties in a complementary way. Through an experiment on a real-dataset, the framework shows to be helpful for explaining complex queries results. Comparing to existing explanation methods, the method could be also considered as an aided-diagnosis method through computing the blame, which helps to understand the impact of uncertain attributes.

A platform that allows the automatic detection of irregular swimming pools is proposed in the next article by Sánchez San Blas et al. The platform uses geographic information tools (GIS) based on orthophotography, combined with advanced machine learning techniques for object detection, as well as a multi-agent architecture, which allows distributed computing and the evaluation of different algorithms combined to improve the detection process. The system has been validated by testing it in different towns in Spain, showing that it is possible to determine the presence of a pool in an image with an accuracy better than 97%.

Focusing on social networks area, the article by Cavaliere et al. presents an emotion-aware solution to analyse users' reactions towards topics constantly discussed over time or in a specific brief period in Twitter. The rationale behind the approach is the combination of emotional analysis of tweet content with the time frequent analysis of relevant topic itemsets, and tweet spread to detect those topics that may have the highest impact. First, the method extracts topics as frequent itemsets from tweets, then the support over time and RoBERTa-based

sentiment analysis are applied to assess the current topic spread and the emotional impact, next a time-grid-based approach allows a granule-level analysis that serves to predict future users' reactions towards topics. Finally, a score function allows building comparative ranked lists of the most relevant topics according to topic sentiment, importance and spread. Experiments demonstrate the potential of the framework on the IEEE COVID-19 Tweets Dataset.

In the same area of social networks, the work by Muñoz et al. develops a methodology for analysing tourism through complex mathematical algorithms, based on unstructured data extracted from social networks. Specifically, graphic and textual data from the profiles of Instagram users feed the classification models. These mine user demographic information and gain insight on what the users were doing in each of their posts, trying to classify that information into any of the categories discovered in the article, acting as a discovery tool for the tourism industry. This has great potential for comparisons on larger amounts of data and even between tourism profiles between cities.

Strukova et al. examine diverse technology-mediated environments that can generate rich data sets through users' interaction and where data can be used to explicitly or implicitly perform a data-driven evaluation of different competencies and capabilities. Their survey revealed four key multimedia environments that include sites for content sharing and consumption, video games, online learning and social networks. The authors found that all these environments are highly correlated with the measurement and development of capabilities such as expertise, language proficiency and soft skills. According to the surveyed studies, this measurement was done with the application of different methods (machine learning, network analysis, natural language processing, statistics and experimental design), which the authors also discuss in detail.

The User Experience Questionnaire (UEQ) is one of very few standard user experience questionnaires available in many different languages. Next article by Hernández-Campos et al. analyses changes in some items of the UEQ for use in the context of Costa Rican culture. Although a Spanish version of the UEQ exists, the authors use a double-translation and reconciliation model for detecting the most appropriate words for Costa Rican culture. These resulted in 7 new items that were added to the original Spanish version. 161 participants participated in a study that examined the original items and the new ones. The results show that the Costa Rican version is neither better nor worse than the original Spanish version, therefore the UEQ is very robust to some changes in the items.

The last article of this issue also relates to user experience area. The objective of the study by Alonso-Virgós et al. is to know if there are "user response" guidelines that a developer with no training or usability experience applies intuitively. Besides the study aims to know the most important recommendations and guidelines, according to the web developers themselves. Knowing the most forgotten recommendations by web developers helps to design effective and specific training in this field.

In closing, I would like to thank both the authors and the reviewers for their commitment to quality assurance and improvement of the articles, aiming for the best reader experience.

Dr. Elena Verdú
Editor-in-Chief

Universidad Internacional de La Rioja

TABLE OF CONTENTS

EDITOR'S NOTE.....	4
WHAT DO WE MEAN BY GENAI? A SYSTEMATIC MAPPING OF THE EVOLUTION, TRENDS, AND TECHNIQUES INVOLVED IN GENERATIVE AI	7
RESEARCH ON BRAIN AND MIND INSPIRED INTELLIGENCE	17
DEEP TRANSFER LEARNING-BASED AUTOMATED IDENTIFICATION OF BIRD SONG.....	33
DEOBFUSCATING <i>LEETSPEAK</i> WITH DEEP LEARNING TO IMPROVE SPAM FILTERING	46
ATTENTIVE FLEXIBLE TRANSLATION EMBEDDING IN TOP-N SPARSE SEQUENTIAL RECOMMENDATIONS	56
SYNTHETIC APERTURE RADAR AUTOMATIC TARGET RECOGNITION BASED ON A SIMPLE ATTENTION MECHANISM	67
POINT CLOUD DEEP LEARNING SOLUTION FOR HAND GESTURE RECOGNITION.....	78
CONVGRU-CNN: SPATIOTEMPORAL DEEP LEARNING FOR REAL-WORLD ANOMALY DETECTION IN VIDEO SURVEILLANCE SYSTEM	88
AN EMPIRICAL EVALUATION OF MACHINE LEARNING TECHNIQUES FOR CROP PREDICTION.....	96
IOT DETECTION SYSTEM FOR MILDEW DISEASE IN ROSES USING NEURAL NETWORKS AND IMAGE ANALYSIS	105
QUANTITATIVE MEASURES FOR MEDICAL FUNDUS AND MAMMOGRAPHY IMAGES ENHANCEMENT ...	117
S-DIVERGENCE-BASED INTERNAL CLUSTERING VALIDATION INDEX	127
EXPLAINING QUERY ANSWERS IN PROBABILISTIC DATABASES	140
A PLATFORM FOR SWIMMING POOL DETECTION AND LEGAL VERIFICATION USING A MULTI-AGENT SYSTEM AND REMOTE IMAGE SENSING.....	153
EMOTION-AWARE MONITORING OF USERS' REACTION WITH A MULTI-PERSPECTIVE ANALYSIS OF LONG- AND SHORT-TERM TOPICS ON TWITTER	166
TOURISM-RELATED PLACENESS FEATURE EXTRACTION FROM SOCIAL MEDIA DATA USING MACHINE LEARNING MODELS	176
A SURVEY ON DATA-DRIVEN EVALUATION OF COMPETENCIES AND CAPABILITIES ACROSS MULTIMEDIA ENVIRONMENTS	182
RESULTS OF A STUDY TO IMPROVE THE SPANISH VERSION OF THE USER EXPERIENCE QUESTIONNAIRE (UEQ)	202
TESTS OF USABILITY GUIDELINES ABOUT RESPONSE TO USER ACTIONS. IMPORTANCE, COMPLIANCE, AND APPLICATION OF THE GUIDELINES	208

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*.

COPYRIGHT NOTICE

Copyright © 2023 UNIR. This work is licensed under a Creative Commons Attribution 3.0 unported License. You are free to make digital or hard copies of part or all of this work, share, link, distribute, remix, transform, and build upon this work, giving the appropriate credit to the Authors and IJIMAI, providing a link to the license and indicating if changes were made. Request permission for any other issue from journal@ijimai.org.

<http://creativecommons.org/licenses/by/3.0/>

What Do We Mean by GenAI? A Systematic Mapping of The Evolution, Trends, and Techniques Involved in Generative AI

Francisco José García-Peñalvo, Andrea Vázquez-Ingelmo *

GRIAL Research Group, Computer Science Department, Universidad de Salamanca, (<https://ror.org/02f40zc51>), Salamanca (Spain)

Received 2 July 2023 | Accepted 20 July 2023 | Published 24 July 2023



ABSTRACT

Artificial Intelligence has become a focal point of interest across various sectors due to its ability to generate creative and realistic outputs. A specific subset, generative artificial intelligence, has seen significant growth, particularly in late 2022. Tools like ChatGPT, Dall-E, or Midjourney have democratized access to Large Language Models, enabling the creation of human-like content. However, the concept 'Generative Artificial Intelligence' lacks a universally accepted definition, leading to potential misunderstandings. While a model that produces any output can be technically seen as generative, the Artificial Intelligent research community often reserves the term for complex models that generate high-quality, human-like material. This paper presents a literature mapping of AI-driven content generation, analyzing 631 solutions published over the last five years to better understand and characterize the Generative Artificial Intelligence landscape. Our findings suggest a dichotomy in the understanding and application of the term "Generative AI". While the broader public often interprets "Generative AI" as AI-driven creation of tangible content, the AI research community mainly discusses generative implementations with an emphasis on the models in use, without explicitly categorizing their work under the term "Generative AI".

KEYWORDS

Artificial Intelligence, Content Generation, Generative AI, Generative Models, Machine Learning, Systematic Literature Mapping.

DOI: 10.9781/ijimai.2023.07.006

I. INTRODUCTION

ARTIFICIAL Intelligence (AI) has evolved as an enthralling topic, attracting the attention of researchers, industry experts, and the general public alike. Its growing popularity may be ascribed to its capacity to produce realistic and creative results and its accessibility, which has far-reaching ramifications in fields such as medicine [1-3], education [4], [5], art [6], [7], music [8], [9], marketing [10], [11], software development [12], [13], among several other areas.

While AI has experienced a surge in popularity in recent years, a particular approach within it has undergone explosive growth during the final months of 2022: the field of generative artificial intelligence or GenAI [14].

The introduction of applications such as ChatGPT¹, Dall-E², or Midjourney³, which make Large Language Models (LLMs) [15], [16] accessible to end-users, has set a milestone in the application of

artificial intelligence to content generation, enabling wide audiences to effortlessly engage in the creation of human-like texts, realistic images, and even music [17].

But what do we exactly mean when we refer to GenAI? What types of content were being generated prior to the emergence of commercial tools like ChatGPT? And for what purposes?

Before diving into the complexities of generative AI, it is crucial to understand the precise meaning and scope of this term, as well as taking a closer look at the content generation processes that existed prior to putting these approaches in the hands of consumers. By investigating these factors, we can shed light on the underlying objectives and motivations driving the adoption of generative AI solutions.

Taking a closer look at the terminology, the word "generative" is defined as "(being) able to produce or create something". If we apply this definition to AI, every model can be technically considered as generative, as they always "produce or create something", whether in the form of numerical predictions or internal rules. However, not every content generation driven by AI is, or has been, considered as Generative AI.

In fact, the term 'Generative AI' has been applied more precisely to models that may produce new, previously unseen information dependent on the data on which they were trained. These models are developing fresh, human-like material that can be engaged with and consumed, rather than just numerical forecasts or internal rules.

¹ <https://chat.openai.com/>

² <https://openai.com/dall-e-2>

³ <https://www.midjourney.com/app/>

* Corresponding author.

E-mail addresses: fgarcia@usal.es (F. J. García-Peñalvo), andreavazquez@usal.es (A. Vázquez-Ingelmo).

The lack of a globally agreed definition of ‘Generative AI’ can result in misunderstanding and misinterpretation. For example, as mentioned before, some may claim that a simple decision tree model that creates rules based on incoming data is a type of Generative AI.

However, the AI research community reserves the term ‘generative’ for more complex models that can create high-quality, human-like material, unlike discriminative models (such as decision tree models), which are trained to predict probabilities of labels given observations [18]. Some examples of the so-called generative models [19] are Generative Adversarial Networks (GANs) or Variational Autoencoders (VAE), among several others (Fig. 1).

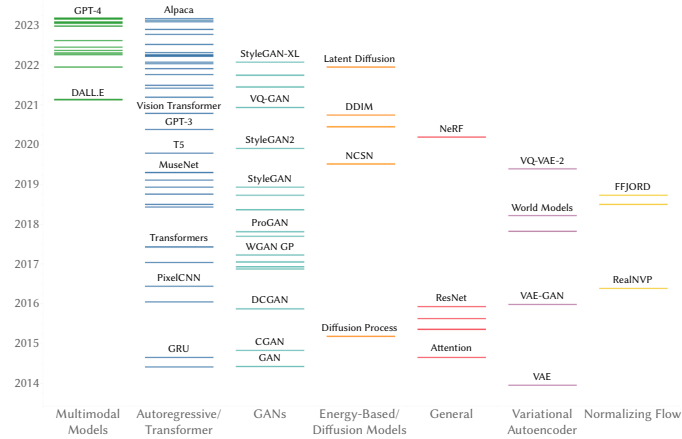


Fig. 1. Timeline of generative models by type. Elaborated by the authors. Source data from [20]. High-resolution version available at <https://doi.org/10.5281/zenodo.8165255>.

But although generative models have already been differentiated from discriminative models by their internal processes and the probabilities they estimate [19], is Generative AI restricted to the use of these kinds of models? Is the underlying model characteristics crucial in affirming that content has been generated through Generative AI? Or do the nature of the content and the ultimate objective hold greater significance?

Without a clear definition, researchers may use the term ‘Generative AI’ to refer to various methods to generate content, leading to confusion and misunderstanding. This can stymie research on this subject, making comparing and contrasting various results difficult.

Furthermore, this ambiguity might make it difficult for industry experts and the general public to comprehend what Generative AI is and what it can achieve, leading to unrealistic expectations or fears about its capabilities, which can affect the adoption and acceptance of this technology.

This work aims at providing an analysis of AI-driven solutions for content generation to further define and delimit the meaning and use cases of Generative AI. This analysis has been carried out through a systematic approach, specifically a systematic literature mapping [21]. A systematic mapping provides a structured framework to thoroughly evaluate existing research in the rapidly evolving field of Generative AI, as well as to identify patterns, and spot potential gaps. It focuses on the extent of the subject rather than its depth, which is crucial in emerging and fast-paced domains. This process aids in establishing clear definitions and boundaries within the field, reducing ambiguity, and fostering a consistent discourse.

The rest of this paper is organized as follows. Section II describes the methodology followed to conduct the systematic mapping, while Section III details the review process. Section IV presents the results obtained from the previous steps. Finally, sections V and VI discuss the results and provide a summary of our main findings.

II. REVIEW PLANNING

This study adheres to the systematic literature review guidelines established by Kitchenham and Charters [22] and the mapping study guidelines set out by Petersen [23], [24]. Specifically, the process is structured around three core stages: planning, conducting, and reporting the findings.

The initial phase involves establishing the primary goal of the review, followed by its development. The main objective of this review is to collect and analyze the existing studies related to the application of AI in content generation, considering the following dimensions: the generated content, the objective of the content generation, the type of models employed and the application domains.

Once the objective has been defined, it is necessary to complete the next two phases, planning and conducting. In these, we define a set of mapping questions (MQs) that will help characterize Generative AI, the inclusion/exclusion criteria, and the search strategy.

A. Mapping Questions

We defined five mapping questions that characterize the AI-driven content generation landscape.

- **MQ1.** How many studies have been published over the years?
- **MQ2.** Who are the most active authors in the area of AI-driven content generation?
- **MQ3.** Which kinds of algorithms and techniques are employed to develop AI-driven content generation applications?
- **MQ4.** Which domains are applying AI-driven content generation to support their studies?
- **MQ5.** What kind of applications were published before and afterwards the popularization of ChatGPT?

These mapping questions are also set to answer the following research question by analyzing the results from the data extraction: **RQ. What do researchers understand by Generative Artificial Intelligence?**

B. Inclusion and Exclusion Criteria

To discard irrelevant works (in terms of the scope of this paper) from the search results, a set of inclusion criteria (IC) and a set of exclusion criteria (EC) are defined, being the **inclusion criteria** as follows:

- **IC1.** The paper’s main objective is the application of content generation (data, images, text, sound, etc.) through artificial intelligence **AND**
- **IC2.** The artificial intelligence solution technical details are identified and described **AND**
- **IC3.** The field in which the solution was applied is identified and described **AND**
- **IC4.** The paper is not a review, survey, or comparative analysis **AND**
- **IC5.** The paper is written in English **AND**
- **IC6.** The paper is published in peer-reviewed Journals, Books, or Conferences **AND**
- **IC7.** The paper is accessible.

The following items refer to the **exclusion criteria** applied:

- **EC1.** The paper’s main objective is not the application of content generation (data, images, text, sound, models, etc.) through artificial intelligence **OR**
- **EC2.** The artificial intelligence solution technical details are not identified nor described **OR**

- **EC3.** The paper is a review, survey, or comparative analysis **OR**
- **EC4.** The field in which the solution was applied is not identified nor described **OR**
- **EC5.** The paper is not written in English **OR**
- **EC6.** The paper is not published in peer-reviewed Journals, Books, or Conferences **OR**
- **EC7.** The paper is not accessible.

These criteria aim at discarding works that are not focusing on generating content through AI. In this sense, we reject studies to benchmark different models, reviews, and works that do not generate tangible content. Following the discriminative and generative models' distinction [18], [19], we want to analyze solutions that generate new data instances in a non-deterministic manner, excluding the outcomes from forecasting, labelling, or classification approaches.

C. Search Strategy

The first step to extracting relevant works for the purpose of this paper is the selection of electronic databases. In this case, two electronic databases are selected: Scopus and Web of Science (WoS). These databases are chosen according to a set of requirements:

- It is a reference database in the research scope.
- It is a relevant database in the research context of this mapping study.
- It allows using similar search strings to the rest of the selected databases and Boolean operators.

An initial search using only the term “generative artificial intelligence” was carried out in these databases. However, this preliminary search yielded a small set of results focused on surveys, editorials, or discussions about the applicability of Generative AI approaches in different domains, such as education.

Given that significant AI-driven content generation applications were not retrieved through this initial search, it was necessary to identify which concepts, approaches or tools are widely associated with Generative AI, to finally collect research literature about these approaches.

Due to the increased accessibility of generative models to consumers after the release of OpenAI's ChatGPT, we analyzed search trends related to “Generative AI” in Google Trends. In this case, we observed that most of the related searches included commercial tools such as “ChatGPT”, “Dall-E”, or “Midjourney”.

Given this trend, we decided to perform another preliminary search, including wildcards to enclose derivations, and the NEAR operator to retrieve works where the terms joined by this operator are separated by an interval of explicitly specified words.

This operator is very handy in this context because we are focused on generative processes driven by AI, so the term related to generation must be near AI-related terms (such as deep learning, machine learning, and so on). This new search was structured as follows:

(“machine learning” OR “deep learning” OR “artificial intelligence” OR “AI” OR “AI-” OR “DL” OR “DL-” OR “ML” OR “ML-”) NEAR/0 (“generat”) OR (“ChatGPT” OR “Midjourney” OR “Dall-E” OR “Dalle” OR “StableDiffusion” OR “Stable Diffusion”) NEAR/1 (“generat*”)*

However, including specific tools would bias the results, as it is nearly impossible to include every released generative AI tool to date. On the other hand, executing a search with the first part of the search string only (*(“machine learning” OR “deep learning” OR “artificial intelligence” OR “AI” OR “AI-” OR “DL” OR “DL-” OR “ML” OR “ML-”) NEAR/0 (“generat*”)*) collected a great set of works, but included an unmanageable quantity of noise, including several articles that were not related to AI-driven content generation.

To overcome these issues, we decided to define further the terminology. As mentioned in section II.B, we want to analyze solutions that generate new data instances in a non-deterministic manner, so we opted to focus on works that explicitly generated content through AI, including images, text, video, audio, sound, etc. We also included terms related to transformations between different types of content, such as text-to-image transformations (e.g., Midjourney and Dall-E).

Once every concept was identified, the specific query strings for each chosen database were specified using their query syntax.

1. Web of Science

TS=(("image generation" OR "text generation" OR "video generation" OR "audio generation" OR "sound generation" OR "3D generation" OR "content generation" OR "code generation" OR "dataset generation" OR "data generation" OR "text to text" OR "text-to-text" OR "text to image" OR "text-to-image" OR "text to audio" OR "text-to-audio" OR "text to video" OR "text-to-video" OR "text to code" OR "text-to-code" OR "text to 3D" OR "text-to-3D" OR "audio to text" OR "audio-to-text") AND ("artificial intelligence" OR AI OR "deep learning" OR "machine learning"))

2. Scopus

TITLE-ABS-KEY(("image generation" OR "text generation" OR "video generation" OR "audio generation" OR "sound generation" OR "3D generation" OR "content generation" OR "code generation" OR "dataset generation" OR "data generation" OR "text to text" OR "text to image" OR "text to audio" OR "text to video" OR "text to code" OR "text to 3D" OR "audio to text" OR "audio to text") AND ("artificial intelligence" OR AI OR "deep learning" OR "machine learning"))

Additionally, we limited the search to **journal articles published over the last 5 years** to analyze recent, established, and complete research works. Using this approach, we collected the final set of articles to analyze and outline the landscape of AI-driven generative solutions.

III. REVIEW PROCESS

The data-gathering process to conduct the present Systematic Literature Mapping has been divided into different phases in which various activities are carried out. The PRISMA 2020 [25], [26] guidelines were followed for data extraction.

Once the search was performed (on May 17th, 2023), the paper selection process was carried out through the following steps:

1. The raw results (i.e., the records obtained from each selected database) were gathered in a GIT repository⁴ and arranged into a spreadsheet. A total of 3295 papers were retrieved: 1835 from Scopus and 1460 from Web of Science.
2. After organizing the records, duplicate works were removed. Specifically, 1332 records were removed, retaining 1963 works (59.58% of the raw records) for the next phase.
3. The maintained papers were analyzed by reading their titles, abstracts, and keywords and by applying the inclusion and exclusion criteria. A total of 1332 papers were discarded as they did not meet the criteria, retaining 631 papers (32.14% of the unique papers retrieved) for the next phase.
4. The selected 631 papers were finally characterized following the mapping questions. For each paper, the following information was collected:
 - a. Content being generated by the AI technique (text, images, code, etc.)
 - b. AI model type employed (transformers, generative adversarial networks, etc.)

⁴ <https://github.com/AndVazquez/slm-gen-ai>

- c. Objective of the AI content generation (data augmentation, image enhancement, text summarization, text translation, style transfer, etc.)
- d. Domain of application of the AI-driven content generation

Fig. 2 shows the PRISMA 2020 [25], [26] flow diagram detailing the data extraction process. The dataset containing the works collected in every phase, along with the 631 selected and characterized works, is available at <https://doi.org/10.5281/zenodo.8162484> [27].

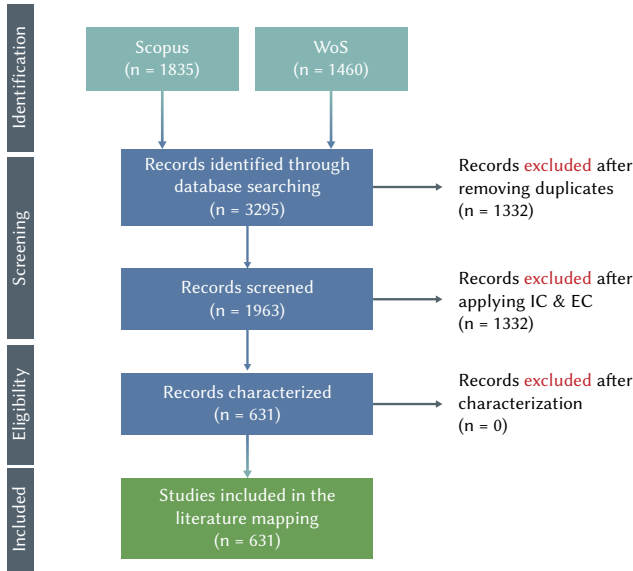


Fig. 2. PRISMA 2020 flow diagram of the literature mapping. High-resolution version available at <https://doi.org/10.5281/zenodo.8167557>.

IV. RESULTS

The following results have been obtained from the analysis of the obtained records. For a comprehensive review of the records, including title, authors, abstract, and characterization, please refer to the “Characterization” sheet of the provided dataset: <https://doi.org/10.5281/zenodo.8162484> [27].

A. How Many Studies Have Been Published Over the Years?

The first mapping question aims at inspecting the temporal landscape of AI-driven content generation. Over the last five years, we can clearly see an increase in the number of works published.

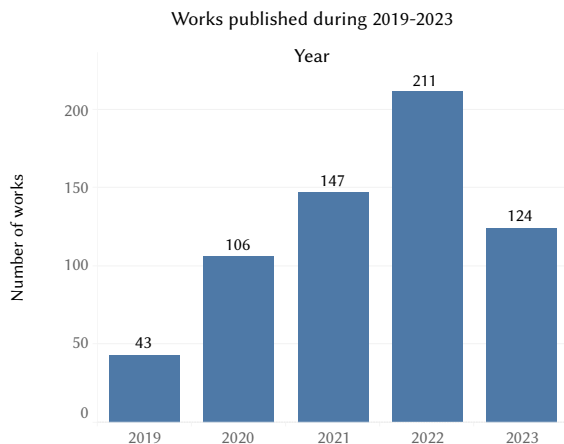


Fig. 3. Number of works published over the last five years. High-resolution version available at <https://doi.org/10.5281/zenodo.8167574>.

It seems that this trend will continue throughout 2023, with 124 articles already published in the first 5 months of the year (Fig. 3).

B. Who Are the Most Active Authors in the Area of AI-driven Content Generation?

We performed an analysis and normalization of the authors involved in the 631 articles retrieved. This analysis allows us to identify influential authors that likely guide the research direction.

TABLE I. MOST PROLIFIC AUTHORS

Articles	Authors
3	Yang, Yang; Chen, Peng; Li, Yibin; Yoon, Hyunsoo; Togo, Ren; Pang, Zhiqi; Ogawa, Takahiro; Haseyama, Miki; Li, Wei; Fujita, Hiroshi; Schlaefer, Alexander; Scarselli, Franco; Fabelo, Himar; Andreini, Paolo; Bianchini, Monica
4	Byun, Yung-Cheol

Most authors (2493) have only published one article in the context of this literature mapping, while 169 have published more than one article. Table I displays the most prolific authors from the records retrieved during the data extraction process.

C. Which Kinds of Algorithms and Techniques Are Employed to Develop AI-driven Content Generation Applications?

As introduced, one of the main concepts of Generative AI is using certain models. But are these models limited to generative models? Or are discriminative models being employed to generate content?

Each article’s primary AI model or technique was identified to answer this question. Fig. 4 shows a clear tendency to use GANs for content generation, followed by encoder-decoder networks (such as Autoencoders) and other types of neural networks. We also found several solutions based on Transformers [28], [29].

Finally, the remaining methods have been grouped under the category “others,” which include hidden Markov methods, evolutionary algorithms, and Bayesian models.

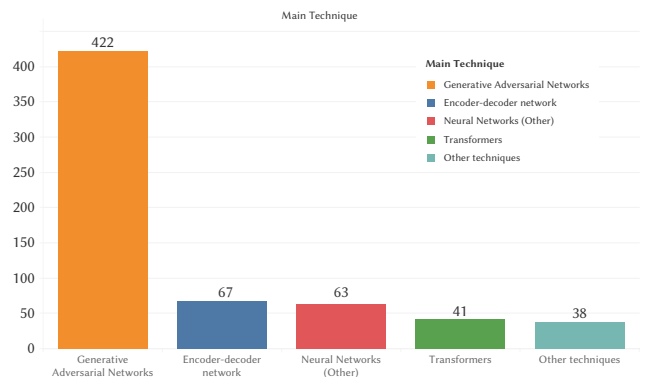


Fig. 4. Number of works grouped by AI technique employed. High-resolution version available at <https://doi.org/10.5281/zenodo.8167582>.

It is important to clarify that although Transformers are encoder-decoder networks, we have decided to include them in a separate category. This separation allows us to analyze the impact of the release of ChatGPT, which is a transformer-based solution, on the usage of this particular model type.

Fig. 5 illustrates the evolution of the number of works utilizing these models. There is a growing trend in adopting Transformer-based solutions, which could be attributed to the recent popularity of GPT models.

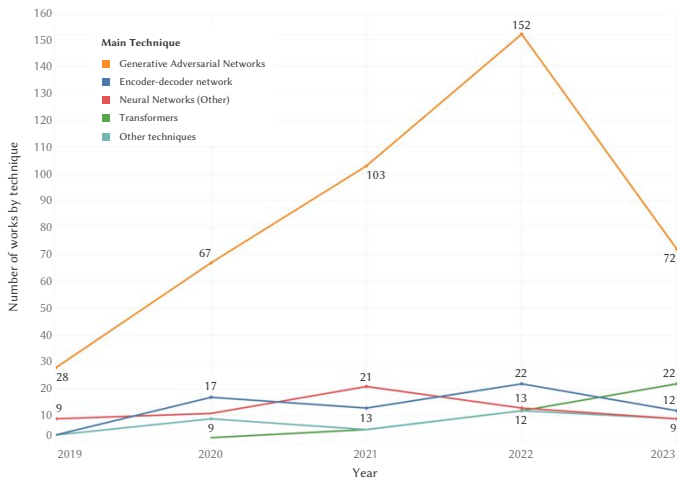


Fig. 5. Number of works grouped by AI technique employed. High-resolution version available at <https://doi.org/10.5281/zenodo.8167600>.

D. Which Domains Are Applying AI-driven Content Generation to Support their Studies?

Another interesting insight regarding generative AI solutions is identifying which domains or fields of study benefit from generative models' outputs.

The main domain of application was extracted from each article. Fig. 6 shows that the principal domains in which generative AI is being applied are medicine and computer vision. Other domains include natural language processing (NLP), machine learning, remote sensing, art, software/videogames development, and cybersecurity.

But for what purposes is AI-driven content generation being used in each domain? This analysis allows us to better understand how and with what objectives generative artificial intelligence is being used in each field of study.

Fig. 7 breaks down the main content generation tasks by domain. It is possible to see that generative AI (and, specifically, Generative Adversarial Networks) is supporting data augmentation in most domains, but especially in medicine. Sample generation provides a minimally intrusive, fast, and effective method to augment or balance

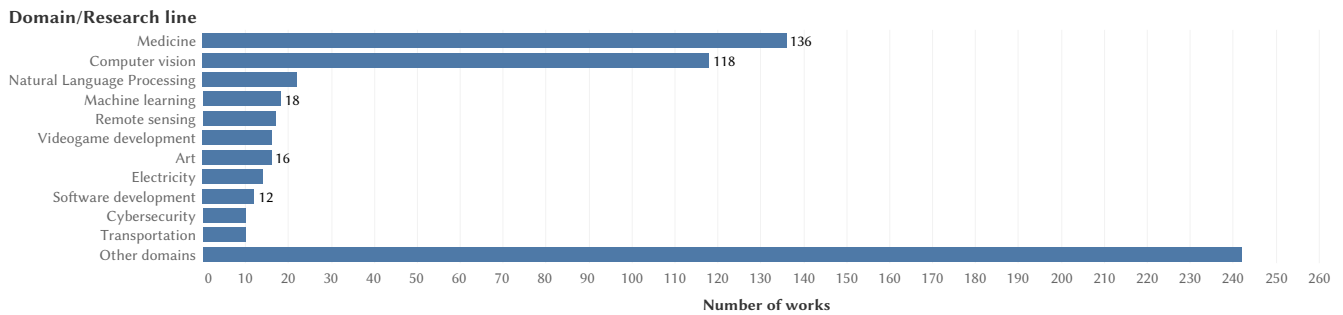


Fig. 6. Number of works grouped by domain of application. High-resolution version available at <https://doi.org/10.5281/zenodo.8167627>.

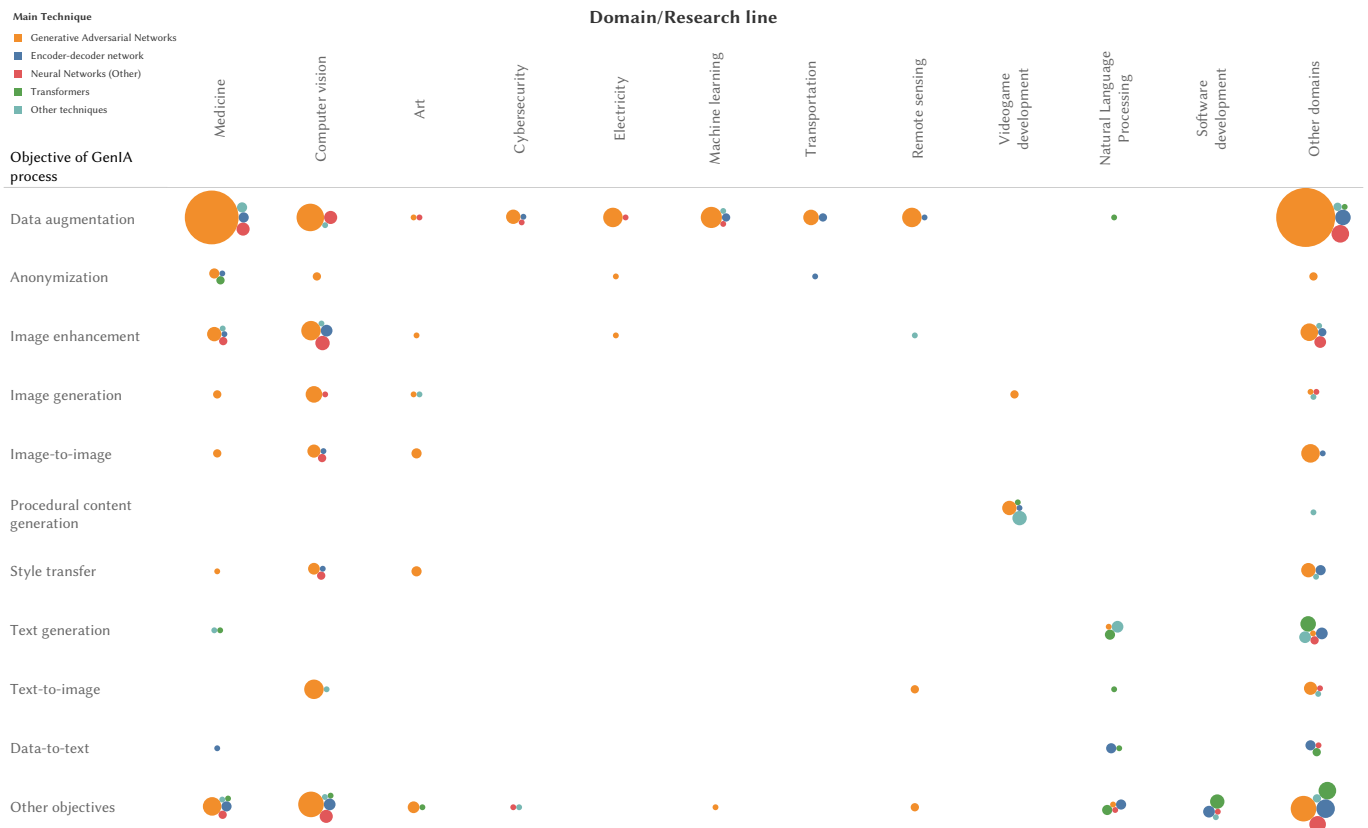


Fig. 7. Number of works grouped by objective, domain and AI technique employed. High-resolution version available at <https://doi.org/10.5281/zenodo.8167632>.

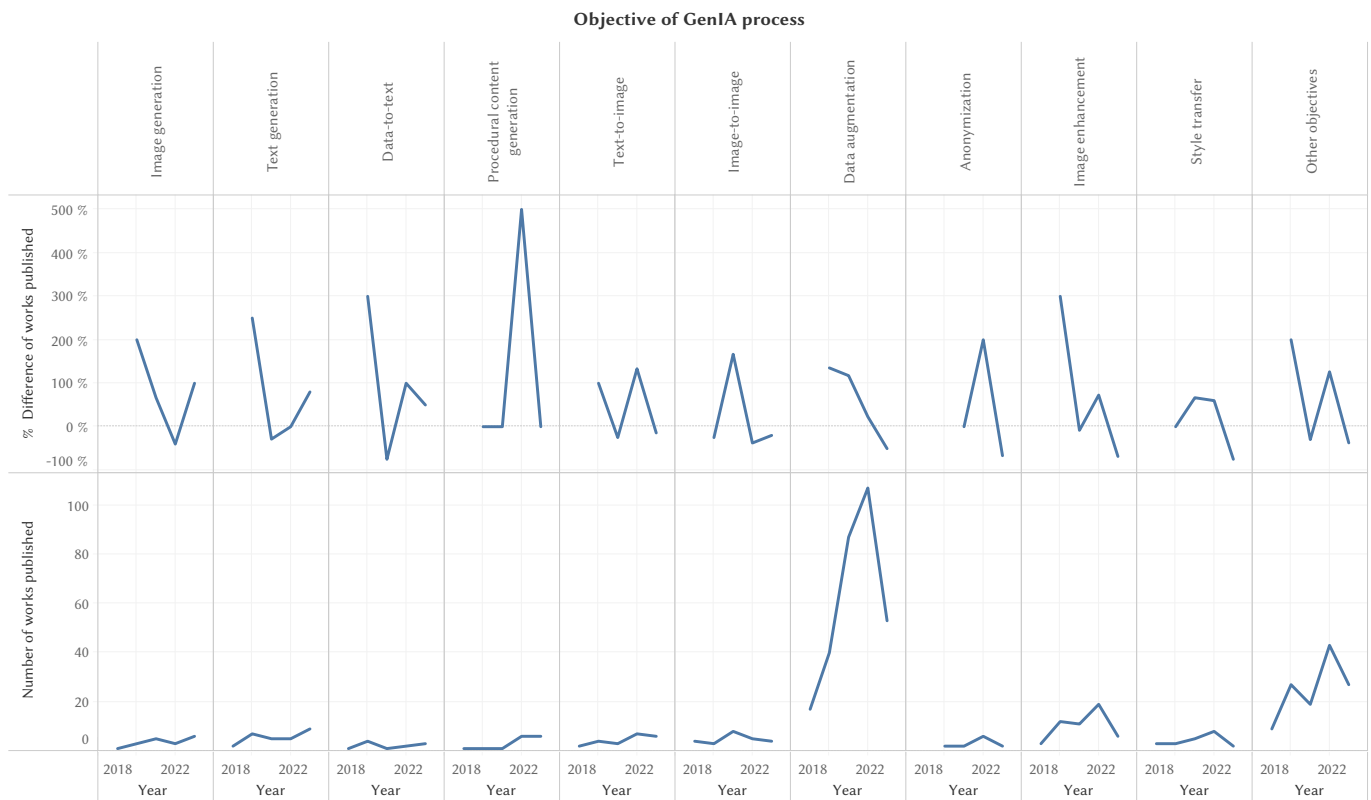


Fig. 8. Difference and number of works published over the years grouped by the objective of the generative process. High-resolution version available at <https://doi.org/10.5281/zenodo.8167647>.

datasets that will subsequently be used to train or fine-tune complex models in each area (e.g., with the aim of diagnosing or segmenting images in the case of medicine).

Another interesting objective of AI-driven content generation is anonymization. By generating new samples, AI can preserve both crucial information and privacy. On the other hand, we also found several objectives related to image processing, including image enhancement (increasing the image resolution, completing missing or damaged parts, or even colorization), style transfer, and text-to-image translations.

Finally, we can also find specific objectives such as procedural content generation in the field of videogame development, and text generation (mostly related to the NLP area).

E. What Kind of Applications Were Published Before and Afterwards the Popularization of ChatGPT?

This question is focused on analyzing the trends in generative AI after the release of ChatGPT⁵, which has significantly reshaped the landscape of AI-driven communication [30], [31]. We have computed the trend of the top tasks supported by AI and compared it over the last five years (Fig. 8).

Although our analysis has covered only the first five months of 2023, we observe a growing trend in text and image generation tasks. This may be influenced by the release of commercial tools for these tasks (ChatGPT, Bing chat⁶, Midjourney, Dall-E, etc.), although it is too early to draw robust conclusions about this.

Additionally, we examined the number of works published annually, segmented by the type of content being generated (such as images, data, text, etc.). It can be observed that images and data

(referring to tabular, geospatial, time-series, or network data) are the types of resources most frequently generated (see Fig. 9).

However, the same trend observed in Fig. 8 is evident for text content. The number of works focused on generating text (spanning activities like human-like text generation, summarization, translation, etc.) has been increasing over the past two years, unlike most other types of generated content analyzed.

Fig. 10 summarizes the main findings of this literature mapping. We can see that Generative Adversarial Networks (GANs) are the preferred generative model in several domains for a wide range of tasks, especially for data augmentation and image-related tasks.

V. DISCUSSION

A. What Do Researchers Understand by Generative Artificial Intelligence (GenAI)?

The concept of “Generative AI” has been in use for several years, but it wasn’t until the mid to late 2010s that it gained widespread recognition. This surge in popularity was in sync with the rise and acceptance of generative models like GANs in the AI research community.

These models, pioneered by Ian Goodfellow and his team in 2014 [32], were crucial in bringing the term “generative AI” into the spotlight.

However, the term gained broader recognition beyond the research community recently. By inspecting Google Trends, we can see that users became interested in this concept around November and December 2022 (Fig. 11), which aligns with the release of ChatGPT and other commercial tools.

Thus, although generative models have existed for several years, the term “generative AI” has only gained popularity with the widespread availability of tools aimed at the general public.

⁵ Launched on November 30, 2022.

⁶ <https://www.bing.com/>

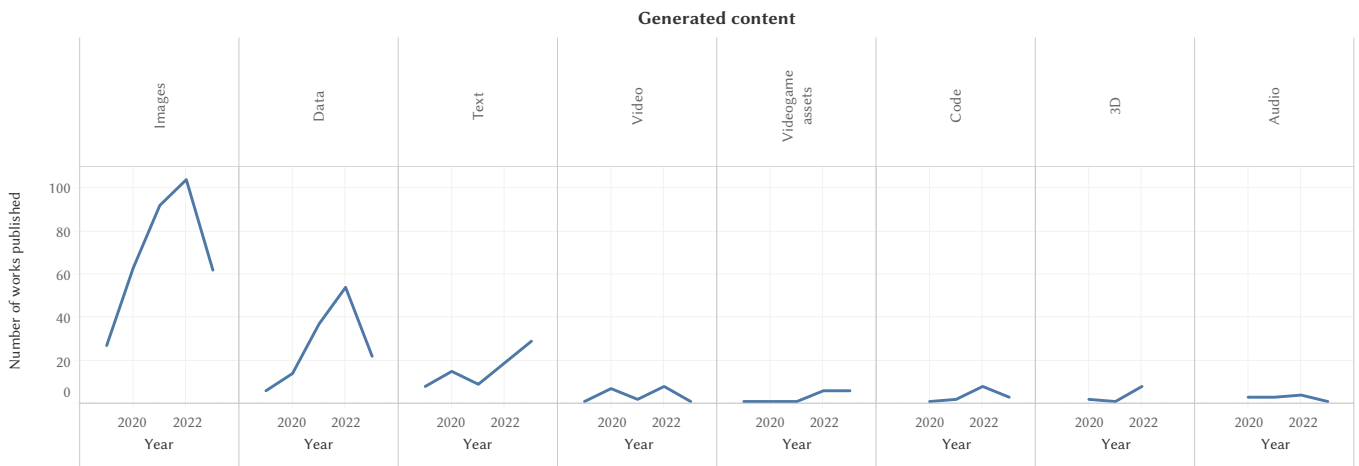


Fig. 9. Number of works published over the years grouped by generated content type. High-resolution version available at <https://doi.org/10.5281/zenodo.8167655>.

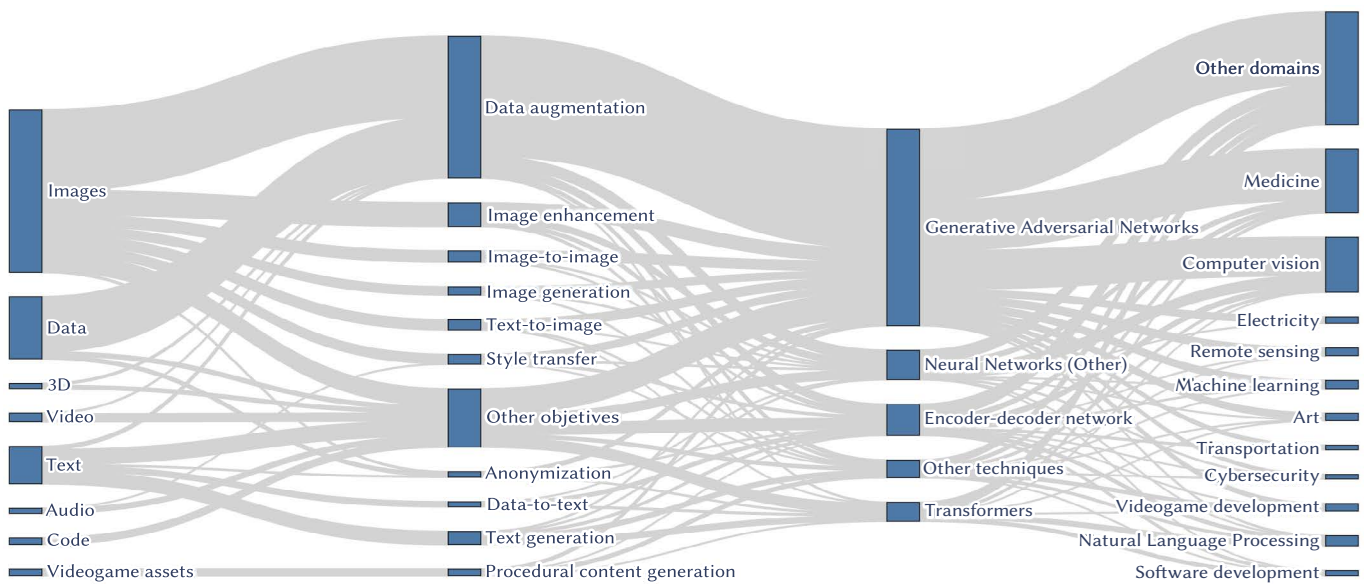


Fig. 10. Relationships among the generated content type, task, AI technique, and application domain in the retrieved works. High-resolution version available at <https://doi.org/10.5281/zenodo.8167662>.

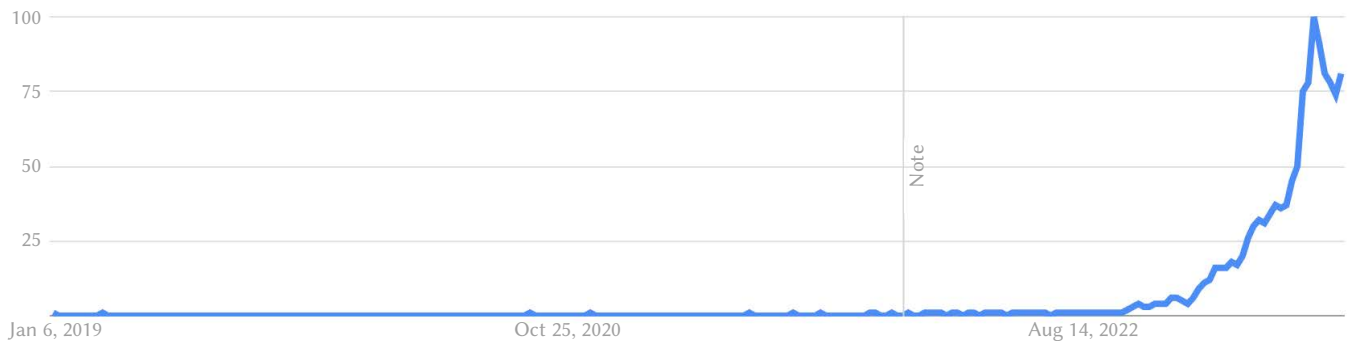


Fig. 11. Worldwide interest over time in the term “Generative AI”. Source: Google Trends, <https://bit.ly/44IzLmI>.

But what does the research community understand by Generative AI? After retrieving and analyzing 631 articles published between January 2019 to May 2023, we obtained a curated set of real-world applications for AI-driven content generation. These solutions included generating a wide variety of resources (images, tabular data, 3D models, videogame assets, etc.) to support different tasks in several domains. What they do have in common is that every solution

employed **generative**, not discriminative models, which could drive the definition of the term Generative AI.

If we further analyze the keywords employed by the researchers to refer to their solutions, only 1 work from 2023 includes the term “Generative AI” (record no. 615 from the “Characterization” sheet of the mapping dataset, <https://doi.org/10.5281/zenodo.8162484> [27]).

In fact, during the preliminary search that we carried out with only words related to Generative AI, we noted that the works referring to this term were mainly editorials or discussions about the implications of commercial tools such as ChatGPT in different domains, but no actual nor applications on generative models producing actual content as we collected in this literature mapping.

So how did the authors refer to the collected solutions? Fig. 12 presents the most common terms found within the abstracts of the 631 retrieved works. We can observe that generation-related terms (generative, generate, generated, etc.) and the generated content (images or data) are very common.

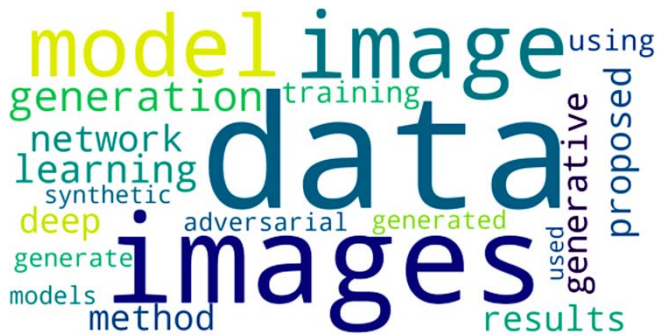


Fig. 12. Most common terms found within the abstracts. High-resolution version available at <https://doi.org/10.5281/zenodo.8167676>.

We also analyzed the most common bigrams within the abstracts to gain more insights into this terminology. Fig. 13 shows how the term “generative” (one of the most common terms found in Fig. 12) was mostly used to refer to generative adversarial networks, the model employed in most works (Fig. 4).

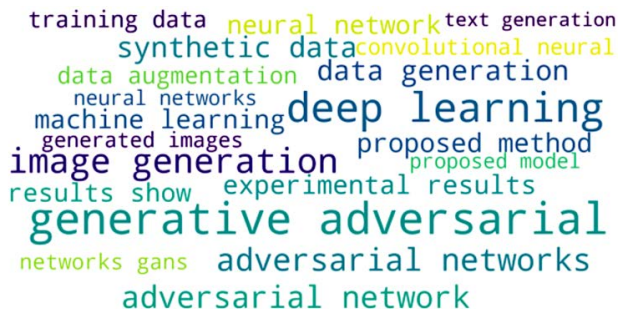


Fig. 13. Most common bigrams found within the abstracts. High-resolution version available at <https://doi.org/10.5281/zenodo.8167693>.

Based on these findings, we can conclude that the general public commonly uses the term “Generative AI” to refer to the creation of tangible content (such as images, text, code, models, audio, etc.) via AI-powered tools. However, the AI research community primarily discusses generative applications focusing on the models used, without explicitly categorizing their work under the term “Generative AI”.

To sum up, following the insights reached through this literature mapping, we can define “Generative AI” as the **production of previously unseen synthetic content, in any form and to support any task, through generative modeling**⁷.

VI. CONCLUSIONS

This work presents the results of a literature mapping about AI-driven content generation. A total of 1963 unique works related to

this topic were analyzed, obtaining 631 categorized articles to better understand the landscape of generative AI solutions. The entire process and final characterization can be reviewed in the provided dataset at <https://doi.org/10.5281/zenodo.8162484> [27].

We found a clear trend in using specific models, such as GANs or encoder-decoder networks, to generate various resources, especially images and tabular data. These solutions have been mostly applied to augment datasets and enhance subsequent models’ predictions.

Although preliminary, it is possible to see how the release of commercial solutions is shifting the landscape of generative solutions, with a slight increase of solutions focused on text generation in the first months of 2023, but also with the advent of new ethical issues and dilemmas, as the widespread accessibility of AI-driven content generation tools has triggered a deep polarization of society regarding Generative AI.

Some individuals are optimistic, envisaging a plethora of opportunities, while others predict dystopian ramifications.

Considering, for example, the domain of education, we can see through the obtained results that Generative AI was marginally applied within this field compared to other areas, such as medicine. However, introducing these tools in education is triggering several concerns among educators, parents, and policymakers [33].

The potential of AI to transform pedagogical methods, assessment systems, and learning experiences opens a new frontier for education. On the one hand, the scalability and personalization offered by AI can improve educational processes by providing a more differentiated and inclusive learning environment.

But just as opportunities and great potential benefits have emerged, there have also grown significant concerns. Some ethical concerns educators raise include assessment, academic integrity [34], and data privacy [35], among others.

The software development field presents a similar narrative. While AI may speed up development processes, automate routine tasks, and significantly reduce debugging time, critics express apprehension about potential job losses and the ethical implications associated with accountability and transparency in AI-generated code [36]. In fact, these powerful applications of AI-driven code generation are also influencing computer science education.

Traditional programming approaches, which frequently rely on manual code writing, debugging, and learning the complexities of programming languages, might be replaced by teaching students how to interface with and manage AI-driven development tools. This change might result in a more efficient learning process, allowing students to handle more complex problems early in their education [37].

However, all these concerns and acceptance issues of Generative AI could be alleviated through a more comprehensive understanding of what precisely Generative AI entails. As introduced, we can technically refer to “Generative AI” as any process of producing any content by any AI technique. But the nuances obtained through this literature mapping offer a deeper view into this term.

By including the technique employed (generative modelling) in the definition of Generative AI, we focus on the process of generating new content from existing resources rather than on the generated content.

It is crucial to understand that Generative AI is not some form of arcane magic but the procedure of training a model with input data. Its capacity to generate original content is based on learning patterns within the available data and then creating outputs that represent these patterns in new ways [18].

By demystifying Generative AI, it is possible to tackle its acceptance issues and address its potential challenges more pragmatically and

⁷ Understood as modeling the joint distribution of inputs and outputs.

effectively. Understanding Generative AI as a data-driven tool rather than an omnipotent solution helps set realistic expectations of what it can accomplish. This viewpoint can facilitate us to successfully integrate AI into different domains without expecting utopian results, minimizing the disappointment that may occur when AI does not perform as anticipated.

ACKNOWLEDGEMENT

This research was partially funded by the Ministry of Science and Innovation through the AvisSA project grant number (PID2020-118345RB-I00).

REFERENCES

- [1] C. Zhang, B. Vinodhini, and B. A. Muthu, "Deep Learning Assisted Medical Insurance Data Analytics With Multimedia System," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 2, pp. 69-80, 2023.
- [2] F. García-Peñalvo *et al.*, "KoopAML: a graphical platform for building machine learning pipelines adapted to health professionals," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. In Press, 2023.
- [3] J. Zhou, T. Li, S. J. Fong, N. Dey, and R. González-Crespo, "Exploring ChatGPT's Potential for Consultation, Recommendations and Report Diagnosis: Gastric Cancer and Gastroscopy Reports' Case," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 2, pp. 7-13, 2023.
- [4] J.-M. Flores-Vivar and F.-J. García-Peñalvo, "Reflections on the ethics, potential, and challenges of artificial intelligence in the framework of quality education (SDG4)," *Comunicar*, vol. 31, no. 74, pp. 37-47, 2023.
- [5] H. Khosravi *et al.*, "Explainable artificial intelligence in education," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100074, 2022.
- [6] A. Chatterjee, "Art in an age of artificial intelligence," *Frontiers in Psychology*, vol. 13, p. 1024449, 2022.
- [7] B. Agüera y Arcas, "Art in the age of machine intelligence," in *Arts*, 2017, vol. 6, no. 4, p. 18: MDPI.
- [8] P. Álvarez, J. García de Quirós, and S. Baldassarri, "RIADA: A Machine-Learning Based Infrastructure for Recognising the Emotions of Spotify Songs," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 2, pp. 168-181, 2023.
- [9] M. Civit, J. Civit-Masot, F. Cuadrado, and M. J. Escalona, "A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends," *Expert Systems with Applications*, p. 118190, 2022.
- [10] 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.
- [11] P. Mikalef, N. Islam, V. Parida, H. Singh, and N. Altwaijry, "Artificial intelligence (AI) competencies for organizational performance: A B2B marketing capabilities perspective," *Journal of Business Research*, vol. 164, p. 113998, 2023.
- [12] R. H. Kulkarni and P. Padmanabham, "Integration of artificial intelligence activities in software development processes and measuring effectiveness of integration," *IET Software*, vol. 11, no. 1, pp. 18-26, 2017.
- [13] A. Mashkoo, T. Menzies, A. Egyed, and R. Ramler, "Artificial intelligence and software engineering: Are we ready?," *Computer*, vol. 55, no. 3, pp. 24-28, 2022.
- [14] T. van der Zant, M. Kouw, and L. Schomaker, "Generative Artificial Intelligence," in *Philosophy and Theory of Artificial Intelligence*, V. C. Müller, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 107-120.
- [15] J. Yang *et al.*, "Harnessing the power of llms in practice: A survey on chatgpt and beyond," *arXiv preprint arXiv:2304.13712*, 2023.
- [16] W. X. Zhao *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.
- [17] F. J. García-Peñalvo, "The perception of Artificial Intelligence in educational contexts after the launch of ChatGPT: Disruption or Panic?," *Education in the Knowledge Society*, vol. 24, art. e31279, 2023.
- [18] D. Foster, "Generative deep learning: teaching machines to paint," *Write, Compose, and Play (Japanese Version)* O'Reilly Media Incorporated, pp. 139-140, 2019.
- [19] H. Gm, M. K. Gourisaria, M. Pandey, and S. S. Rautaray, "A comprehensive survey and analysis of generative models in machine learning," *Computer Science Review*, vol. 38, p. 100285, 2020/11/01/ 2020.
- [20] D. Foster. (2023). *Generative Deep Learning - 2nd Edition* Codebase. Available: https://github.com/davidADSP/Generative_Deep_Learning_2nd_Edition/tree/main
- [21] F. J. García-Peñalvo, "Developing robust state-of-the-art reports: Systematic Literature Reviews," *Education in the Knowledge Society*, vol. 23, p. e28600, 2022.
- [22] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering. Version 2.3," School of Computer Science and Mathematics, Keele University and Department of Computer Science, University of Durham, Technical Report EBSE-2007-01, 2007, Available: <https://goo.gl/L1VHcw>.
- [23] K. Petersen, S. Vakkalanka, and L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Information and software technology*, vol. 64, pp. 1-18, 2015.
- [24] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12*, 2008, pp. 1-10.
- [25] M. J. Page *et al.*, "PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews," *BMJ*, vol. 372, p. n160, 2021.
- [26] M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021.
- [27] A. Vázquez-Ingelmo and F. J. García-Peñalvo, "Dataset for the mapping study 'What do we mean by GenAI?'," (1.0) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.8162484>, 2023.
- [28] R. Gruetzemacher and D. Paradice, "Deep transfer learning & beyond: Transformer language models in information systems research," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1-35, 2022.
- [29] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017, vol. 30, pp. 5998-6008.
- [30] A. S. George and A. H. George, "A review of ChatGPT AI's impact on several business sectors," *Partners Universal International Innovation Journal*, vol. 1, no. 1, pp. 9-23, 2023.
- [31] A. Bozkurt *et al.*, "Speculative Futures on ChatGPT and Generative Artificial Intelligence (AI): A collective reflection from the educational landscape," *Asian Journal of Distance Education*, p. (Published), 2023.
- [32] I. Goodfellow *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.
- [33] F. J. García-Peñalvo, F. Llorens-Largo, and J. Vidal, "The new reality of education in the face of advances in generative artificial intelligence," *RIED: Revista Iberoamericana de Educación a Distancia*, vol. 27, 2023.
- [34] W. M. Lim, A. Gunasekara, J. L. Pallant, J. I. Pallant, and E. Pechenkina, "Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators," *The International Journal of Management Education*, vol. 21, no. 2, p. 100790, 2023.
- [35] L. Tredinnick and C. Laybats, "The dangers of generative artificial intelligence," ed: SAGE Publications Sage UK: London, England, 2023, p. 02663821231183756.
- [36] B. Meyer. (2022). *What Do ChatGPT and AI-based Automatic Program Generation Mean for the Future of Software*. Available: <https://bit.ly/3LyAJLj>
- [37] O. Hazzan. (2023). *ChatGPT in Computer Science Education*. Available: <http://bit.ly/3WYTxpx>



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 her bachelor's degree in Computer Science from the University of Salamanca (USAL), in 2016, her master's degree in Computer Science (USAL) in 2018, and her Ph.D. degree in Computer Science (USAL) in 2022. She is a member of the Research Group of Interaction and eLearning (GRIAL) since 2016. Her area of research is related to human-computer interaction, software engineering, data visualization, and machine learning applications.

Research on Brain and Mind Inspired Intelligence

Yang Liu¹, Jianshe Wei² *

¹ Henan Key Laboratory of Big Data Analysis and Processing, College of Computer and Information Engineering, Henan University, Kaifeng (China)

² Laboratory of Brain Function and Molecular Neurodegeneration, Institute for Brain Science Research, School of Life Sciences, Henan University, Kaifeng (China)

Received 7 August 2022 | Accepted 4 June 2023 | Published 13 July 2023



ABSTRACT

To address the problems of scientific theory, common technology and engineering application of multimedia and multimodal information computing, this paper is focused on the theoretical model, algorithm framework, and system architecture of brain and mind inspired intelligence (BMI) based on the structure mechanism simulation of the nervous system, the function architecture emulation of the cognitive system and the complex behavior imitation of the natural system. Based on information theory, system theory, cybernetics and bionics, we define related concept and hypothesis of brain and mind inspired computing (BMC) and design a model and framework for frontier BMI theory. Research shows that BMC can effectively improve the performance of semantic processing of multimedia and cross-modal information, such as target detection, classification and recognition. Based on the brain mechanism and mind architecture, a semantic-oriented multimedia neural, cognitive computing model is designed for multimedia semantic computing. Then a hierarchical cross-modal cognitive neural computing framework is proposed for cross-modal information processing. Furthermore, a cross-modal neural, cognitive computing architecture is presented for remote sensing intelligent information extraction platform and unmanned autonomous system.

KEYWORDS

Brain and Mind Inspired Intelligence, Brain and Mind Inspired Computing, Cognitive Computing, Cross-Modal Cognitive Neural Computing, Deep Learning, Multimedia Neural Cognitive Computing.

DOI: 10.9781/ijimai.2023.07.004

I. INTRODUCTION

BRAIN and Mind inspired Intelligence (BMI) is an innovative bio-inspired computing based on bionics, which enlightened by cognitive function, neural structure and system behavior. BMI would realize state-of-the-art intelligence system which has advanced in computing ability, efficiency, and energy consumption. BMI will establish fundamental theories and models of neural cognitive computing, explore the algorithm and technology of the new generation computation, and research architecture and system of Artificial General Intelligence (AGI). The research contents of BMI include the intelligence scientific theory, brain-inspired algorithms and brain-like hardware for learning and processing.

Although BMI has been an obvious success at present, it is far from reaching the general autonomous intelligence level, and it lacks the ability of multimedia perception and cross-modal cognitive in both models and algorithms. For the study of the brain-inspired algorithm, which one is brain-inspired cognitive simulation from the global macroscopic functional, the other is reverse engineering the brain-like [1] neural structural emulation from the microscopic structures of neurons, synapses, and networks. However, there is still a lack of effective research how to assemble advanced function of the complex system from the local network in mesoscopic. There is still a long way to go to study the gap between natural intelligence and brain-

inspired intelligence [2]. Until now, the study of the brain-inspired model has not supported the uniform mind function such as sensation, perception, cognition and behavior.

A. Key Contribution of the Paper

To address the problems of scientific theory, common technology and engineering application of multimedia and multimodal information computing, this paper is focused on the theoretical model, algorithm framework and system architecture of BMI based on the structure mechanism simulation of nervous system, the function architecture emulation of cognitive system and the complex behavior imitation of natural system.

- (1) We reviewed the current state, key trends and outstanding issues of BMI from brain research projects, mind research projects, statistical learning, cognitive computing, deep learning and neuromorphic computing.
- (2) Based on information theory, system theory, cybernetics and bionics, the related concept of brain and mind inspired computing (BMC) are defined, and the hypothesis, model and framework of the frontier BMI theory are proposed.
- (3) A semantic-oriented Multimedia Neural Cognitive Computing (MNCC) model is designed for multimedia semantic computing based on the brain mechanism and mind architecture.
- (4) A hierarchical Cross-modal Cognitive Neural Computing (CCNC) framework is proposed for cross-modal information processing.
- (5) Furthermore, a Cross-modal Neural Cognitive Computing (CNCC) architecture is presented for remote sensing intelligent information extraction platform and unmanned autonomous system.

* Corresponding author.

E-mail addresses: jswei@henu.edu.cn (J. Wei), ly.sci.art@gmail.com (Y. Liu).

The rest of the work is structured as: Section II provides a research background, which deals with a literature review, including studies of BMI, brain and mind research project, statistical learning, cognitive computing, deep learning and neuromorphic computing. Section III studies the relationship between the structure mechanism of the nervous system and the function architecture of the cognitive system, and then gives the definition of the BMI related concepts. Section IV gives the formal description and algorithm design of MNCC model and CCNC framework, and then gives the application results of engineering system based on CNCC architecture. Finally, we conclude our work and propose potential future work on this topic in Section V.

II. RELATED WORK

A. Research Projects of Brain and Mind

With the development of Turing machine proposed in 1936 and the birth of first electronic computer ENIAC in 1946, the information technology has lay a foundation for realizing the dream of intelligent technology. However, since the concept of artificial intelligence was proposed in 1956, the related research fluctuates between success and failure. Many countries have made the significant investment in the scientific research of artificial intelligence in the past 66 years. Some initiatives and projects about intelligent behavior, brain and mind have been proposed respectively, such as Decade of the Brain (1990-2000), Decade of the Behavior (2000-2010) and Decade of the Mind (2012-2022). In addition, brain research projects of some countries have been launched successively, such as US BRAIN Initiative, EU Human Brain Project, Japan Brain/MINDS, China Brain Project, Brain Canada, Australian Brain Initiative, Korea Brain Initiative and Israel Brain Technologies, etc.

The neuroscience methods for studying the brain and the cognitive science methods for studying the brain tend to fuse and interact. Elucidation of the neural mechanisms in brain and cognitive processes in mind allows us to understand mind principles, and facilitates the intervention and diagnosis of neurological and psychological diseases. It also contributes to the research of frontier scientific theories, the development of key technologies and the application of engineering systems of BMI, and provides the basis for a next generation artificial intelligence with design beyond von Neumann architecture.

B. Cognitive Computing

The Bayesian theory has an indispensable role in the statistical learning. The Bayesian mechanism of the brain has been validated by a lot of experimental results in psychology and neurophysiology. According to Bayesian probability, causal inference, and statistical theory, it can simulate the perception and cognitive process of visual and aural, which can construct a unified cognitive theoretical framework.

As a successful technique and powerful method, cognitive computing has existed for a long time, but it has been making a breakthrough in recent years. Literature [3] seeks nothing less than to discover, demonstrate, and deliver the core algorithms of the macaque monkey brain. Watson system based on DeepQA and transfer learning to simulate cognitive processes of mind such as learning, thinking and decision making.

However, traditional cognitive computing focuses on the simulation of mind function, and lacks in-depth research on the emulation of brain structure mechanism, which makes it difficult to realize general artificial intelligence. In addition, the role of belief in automatic reasoning is worth exploring [4].

C. Neuromorphic Computing

There are two main ways to build general artificial intelligence, namely Brain-Inspired Computing (BIC) and Brain-Like Computing (BLC). BIC simulates and designs intelligent model inspired by top-down brain function, including Artificial Neural Network (ANN) and deep learning. BLC emulates bottom-up brain structure to realize intelligent system. NeuroMorphic Technologie (NMT) includes three main forms: NeuroMorphic Engineering (NME) of neurons is established by sub-threshold analog circuit, NeuroMorphic Computing (NMC) [5] of spiking neural network is realized by digital system, and NeuroMorphic Device (NMD) of spiking neural network memristor-based [6] is constructed with new memory materials.

Carver Mead envisioned build neuromorphic electronic systems based on analog VLSI (very large scale integration) circuits, which established a new paradigm in hardware computing of BLC [7]. Guided by brain-like 'spiking' computational frameworks, NMC is modeling and emulation of the bionic brain for machine intelligence-promises to realize artificial intelligence while reducing the energy requirements of computing platforms [8]. NMC mimics neuro-biological architectures by VLSI systems containing electronic analog circuits, which aiming at brain-like capabilities and efficiencies.

Spiking Neural Networks (SNN) is also known as the third generation of neural network models, which increases the level of realism in a neural emulation. Spiking neurons model includes Hodgkin-Huxley, Leaky Integrate and Fire (LIF), Spike Response Model (SRM) and Izhikevich etc. Besides neuronal and synaptic state, SNN also incorporated the concept of time into their operating model. The SNNs exploit spatio-temporal information based on sparse and dynamic spiking event, and have advantage of low-power computing. The spiking neurons have a discontinuous activation function, and emit discrete spikes that are nondifferentiable; hence it cannot use directly the gradient-descent BackPropagation (BP) algorithm to training SNNs. Currently, the existing training methods for SNNs fall into three types: (1) unsupervised learning, such as spike-timing-dependent plasticity (STDP), Hebbian learning; (2) supervised learning, such as SpikeProp, Remote Supervised Method (ReSuMe), FreqProp, ANN-to-SNN conversion, Spatio-Temporal BackPropagation (STBP) [9] and Quantum Superposition SNN (QS-SNN) [10]; (3) reinforcement learning, such as Spiking Actor-Critic(SAC) method, reinforcement learning through reward-modulated STDP.

At present, BMI has achieved remarkable achievements in the brain-like neuronorphic technology. A large number of 'Big Brain' chips and systems have been designed, such as NeuroGrid [11], BrainScaleS [12], SpiNNaker [13], Darwin NPU [14], Tianjic [15], TrueNorth [16], Memristor, TPU [17], Loihi, DianNao [18] family, DishBrain [19], and DeepSouth [20]. In addition, synaptic efficacy and synaptic plasticity can be accomplished using emerging non-volatile memristive technologies such as resistive random-access memory (RRAM), phase-change memory (PCM), floating-gate transistor, and memristive dot products. Furthermore, the much brain-inspired software system has been developed, such as SpikeNET, NEURON, GENESIS, BRAIN and NEST emulator etc. Compass is a multi-threaded; massively parallel functional emulator and a parallel compiler that mapping a network of Long Distance Pathways (LDP) in the brain to TrueNorth [21]. Literature [22] present neuron models of the brain as Spaun (semantic pointer architecture unified network) for Nengo, which can emulate the human tasks, such as image recognition, serial working memory, reinforcement learning, counting, question answering, rapid variable creation, and fluid reasoning.

However, as a key technology of BLC, NMT needs to be studied on the basis of understanding the structure and neural mechanism of the brain. Before the mechanism of neural system is not completely clear,

it is very difficult to realize NMT in algorithm principle and model design. NMT inspired by the brain promise fundamentally different ways to process information with extreme energy efficiency and the ability to handle the avalanche of unstructured and noisy data. So brain-inspired computing needs a master plan [23].

D. Deep Learning

Deep learning, also known as feature learning, benefits from the combination of big data and high-performance computing. The development of deep learning has gone through four stages: (1) McCulloch and Pitts put forward a neuron model for logic and computation in 1943, which created the first Computational Theory of Mind and Brain (CTMB) [24]; (2) Rosenblatt proposed perceptron for linear classification in 1958; (3) Multi-layer perceptron (MLP) based on BP algorithm is used to solve nonlinear problems; (4) Hinton proposed deep learning in 2006.

In essence, deep learning is also a classic neural computing of structuralist technology. Combined with the traditional machine learning algorithms of functionalist and behaviorist, it can effectively solve the problems of high energy consumption and low intelligence of the existing computing system. For example, AlphaGo improved the performance of the Go program with CNN, reinforcement learning, and Monte Carlo tree search algorithm [25],[26] It is mathematically equivalent to using an MLP after the convolutional layers. With the application of deep learning in engineering systems, cognitive mechanisms such as perception, attention, memory and emotion are widely used in the intelligent processing of multimedia such as image, video, audio and natural language. The more and more novel Deep Neural Network (DNN) were designed, such as Forward-Forward (FF) algorithm and capsule network [27], Generative Adversarial Network (GAN) [28]. For example, it can greatly improve the performance of machine translation and intelligent retrieval in natural language processing by the employed Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) [29], DeepFair [30], Encoder-Decoder model, Transformer [31] and GPT-3 model with 175 billion parameters [32]. DNN is also a graph model in nature. Inspired by the great success of deep learning in machine learning tasks which are typically represented in the Euclidean space, Graph Neural Network (GNN) [33] is also introduced to solve the learning task of non-Euclidean domains.

However, as the most popular technology of BIC, deep learning is a black box model that lacks explainability. There are also technical bottlenecks in small samples and energy consumption, and it cannot be applied in the environment with high security requirements.

III. SCIENTIFIC THEORY BASED ON BRAIN AND MIND INSPIRED INTELLIGENCE

For the study of brain and mind, semantic computing is a complex scientific problem in a visual or auditory scene. It has an essential enlighten to realize semantic computing of target recognition, and multimedia intelligent processing that the function, structure and social behavior of the cognitive framework and the neural mechanism. With the rapid engineering development of deep learning and cognitive computing in the field of artificial intelligence, more and more heuristic algorithms based on biological intelligence have emerged. However, there are essentially different from scientific theory research, technical route and method, and implementation of engineering and system among neuroscience, cognitive science and computing science. In order to solve the complex problem of audio-visual semantic computation, the computational model is established urgently for mimicking the brain and mind, inspired by the framework of cognitive function and the mechanism of neural processing.

A. Definition of Related Concepts of Brain and Mind Inspired Intelligence

Currently, there are two research directions which have attracted much attention in BMI. One is the Brain Inspired Computing (BIC) method based on systematic behavior to simulate cognitive function, the other is the Brain Like Computing (BLC) method based on neuron, synapse or local network structure to emulate neural mechanism. Considering that brain inspired intelligence, BIC and BLC have no exact definitions at present, these terms may be confused with each other. This section explains their differences and strictly defines the concepts related to Brain and Mind inspired Intelligence Theory (BMI Theory).

Definition 1 (BII Theory). Brain Inspired Intelligence (BII) is an intelligent model, method and system that enlightened human brain and realizes brain inspired intelligence information processing.

Definition 2 (BIC Technology). Brain Inspired Computing (BIC), also known as mind inspired computing, is a computing method inspired by the principles of cognitive mechanisms and mental functions, which simulates mental models from the functional level, and explores and designs mind inspired intelligence algorithms.

Definition 3 (BLC Technology). Brain Like Computing (BLC) is a network architecture that emulates the brain nervous system. It focuses on the emulation of the nervous system at the system structure level, and explores and designs brain like intelligence algorithms.

Definition 4 (BMI Theory). Brain and Mind inspired Intelligence (BMI) is an innovative bio-inspired computing intelligent scientific theory, intelligent technology and intelligent engineering based on bionics. The inspiration for establishing BMI theory comes from functional architecture of cognitive system, structural mechanism of nervous system, and complex behavior of natural system.

Definition 5 (BMC Technology). Brain and Mind inspired Computing (BMC) is a intelligent model, intelligent algorithm and intelligent system that mimicking brain structure, mind functional and human behavior at the same time. BMC researches intelligent theoretical model and algorithm of perceptual computing and cognitive computing, and realizes new generation brain inspired and mind like intelligent system.

BMC focuses on the structural simulation of nervous system, the functional emulation of cognitive system and the behavior imitation of natural intelligent system, so as to solve the compatibility of functional emulation and structural simulation of intelligent computing. Firstly, BMC emulates the network structure of nervous system and brain information processing architecture. Secondly, BMC simulates the functional principle and mental information processing model of cognitive system, Thirdly, BMC imitates the complex system behavior of natural intelligence.

Aiming at the unstructured and complex semantic processing of multimedia computing, we propose a multimedia neural cognitive computing model.

Definition 6 (MNCC Model). Multimedia Neural Cognitive Computing (MNCC) is to construct multimedia information processing model and algorithm for the purpose to solve the problems of semantic processing of unstructured, massive, multi-modal, and temporal-spatial distribution in multimedia information. MNCC is a BMC model, which establishes a new generation of the multimedia information processing model and algorithms with cognitive computing of system behavior in macroscopic level, and neural computing of physiological mechanisms in microscopic level.

MNCC focuses on the perceptual computation and semantic cognitive processing of multimedia interactive information, especially on the feature extraction, content analysis and semantic computation

of visual media, auditory media, natural language and other media information. MNCC is a BMI model for multimedia intelligent processing based on the structure and mechanism of the nervous system, and the function and framework of the cognitive system.

In order to realize cross-modal content processing and solve the audio-visual cross-media computing problem, we further propose a cross-modal cognitive neural computing framework.

Definition 7 (CCNC Framework). Cross-modal Cognitive Neural Computing (CCNC) is cross-modal information processing framework and mind inspired cognitive computing method of cross-media knowledge reasoning based on MNCC model. CCNC is a BMC framework, which mainly solves the problems of cross-modal semantic computing by the mechanism of multisensory integration and multimodal cooperative cognitive.

CCNC focuses on the information fusion of different perception media, and multimedia semantic correlation. It usually explores cross-media computing methods based on temporal and spatial correlation. To overcome the bottleneck of cross-modal semantic computing, CCNC researches hierarchical framework and algorithm based on mind-inspired cognitive computing and brain-inspired deep learning.

Aiming at the problems of cross-modal target recognition in remote sensing intelligent information extraction platform, and cross-media semantic recognition in unmanned autonomous system, we further propose a cross-modal neural cognitive computing architecture.

Definition 8 (CNCC Architecture). Cross-modal Neural Cognitive Computing (CNCC) is a kind of system architecture of brain inspired cross-modal intelligence. On the basis of MNCC model and CCNC framework, CNCC adopts transfer learning, multi-modal cooperative cognitive mechanism of neural information and modal mutual information to design the information processing system architecture of cross-modal brain inspired intelligence at the level of cognitive computing, so as to solve the complex engineering application problems of cross-modal semantic computing and transfer learning.

CNCC architecture focuses on system engineering implementation of the media semantics cognition and cross-modal correlation. CNCC can be applied to cross-modal target recognition with different modal perception information, such as visible light, infrared, radar, sonar, etc.

B. Structure and Mechanism of the Nervous System

It are the source of bionic inspiration for the study of BMI theory that the structure and mechanism of the nervous system, and function and the framework of the cognitive system. There is a systematic study of the brain's information processing mechanism in different disciplines. On the one hand, the neuroscience analysis brain mechanism of neural processing at the levels of the cortical structure and the neural circuits based on the white box method. On the other hand, cognitive science research model of mind's information processing through cognitive function and the phenomenon based on the black box method. However, artificial intelligence of computer science realizes logic computation of a finite state machine based on Turing machine and Von Neumann architecture. Although computer science has made great advances, the principle, structure and function of brain, mind and computer are essentially different.

Behind the neurodynamic characteristics of brain structure and the emergence of social behavior of mental function, there are the basic laws of complex intelligent system. Generally, the neocortex of the cerebellum is the core component of intelligent processing in the field of neurocognitive science; the thalamus is the switch of information entry and selective attention; the limbic system and the hippocampus are the controllers for memory and emotion. The human central nervous system is composed of white matter and grey matter, and has

the obvious symmetry and contralateral. The neocortex structure of the gray matter is similar to the digital-analog electronic processing unit with processing linear and nonlinear function. The LDP of the white matter made up the complex White Matter Network (WMN). It can be regarded as wiring diagram of neural processing. So the neocortex structure and WMN are very important for understanding the overall structure of the brain. The research on the computable model of nervous system includes Cortical Model (see Appendix A), LDPs and Neural Circuits (see Appendix B). Among them, Hierarchical Temporal Memory(HTM) [34] is very noteworthy (see Appendix C).

Fig. 1 is our WMN visualization of 48 brain regions from the high accuracy human brain LDP database in documents [35],[36], which consider WMN connection weights and nodes size. The analysis found that there were a large number of long-distance loops in the human brain WMN, and there were also a large number of connections between the thalamus and cortex. So the relay nuclei of the thalamus are a controller of the selective attention to sensory information, and the association nuclei of the thalamus are a switch of cortical processing information.

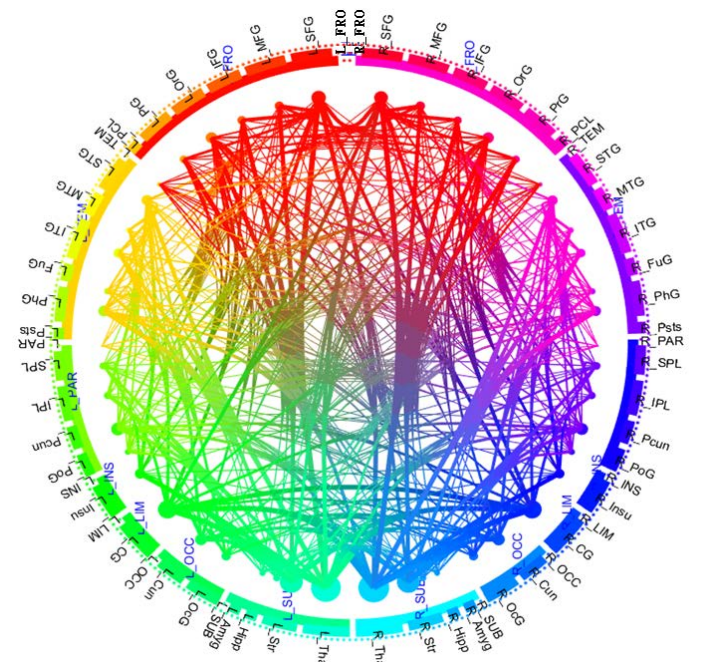


Fig. 1. WMN structures with connection weights visualization on 48 brain regions.

The hierarchy of cortical function and structure of neural pathways and circuits can provide significant evidence for the neural cognitive model. Neural circuits are the important material of relevance such as feedback, stochastic resonance, recurrence iterative, resonance, memory, emotion, attention, language, and thinking. It can inspire for build neural model that the hierarchical structure of cortical function, and the structure between the neural pathways and loops. Studies indicate that the central nervous system is a scale-free and small-world complex network. On the basis of the WMN visualization and document [37],[38], we propose a simplified structure model of the whole brain WMN in Fig. 2.

C. Function and Architecture of the Cognitive System

Generally, the human's mind activities involve many aspects in cognitive neuroscience. Specifically, it includes sensation (such as light, sound, touch, taste, smell, etc.), perception (such as seeing, hearing, feel, tasting, smelling, etc.), behavior (such as movement, reaction, choice, interaction, etc.), and cognition (such as attention, memory, emotion, logic, language, reasoning, understanding, problem-

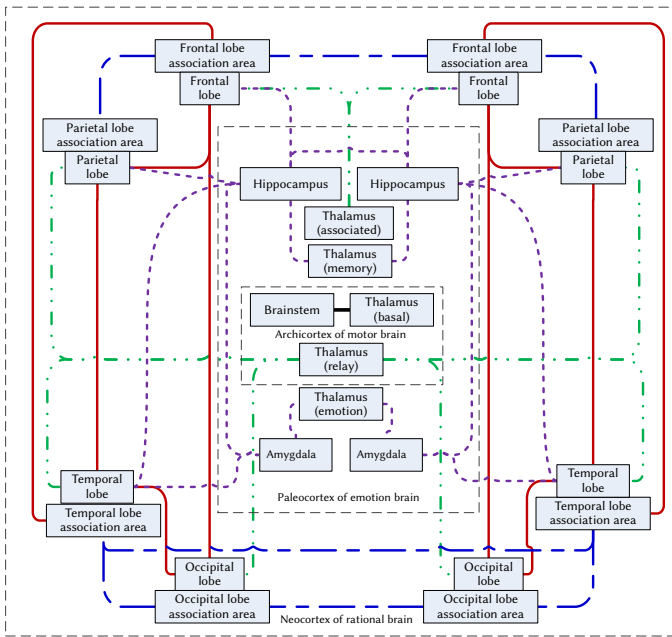


Fig. 2. The WMN structure model.

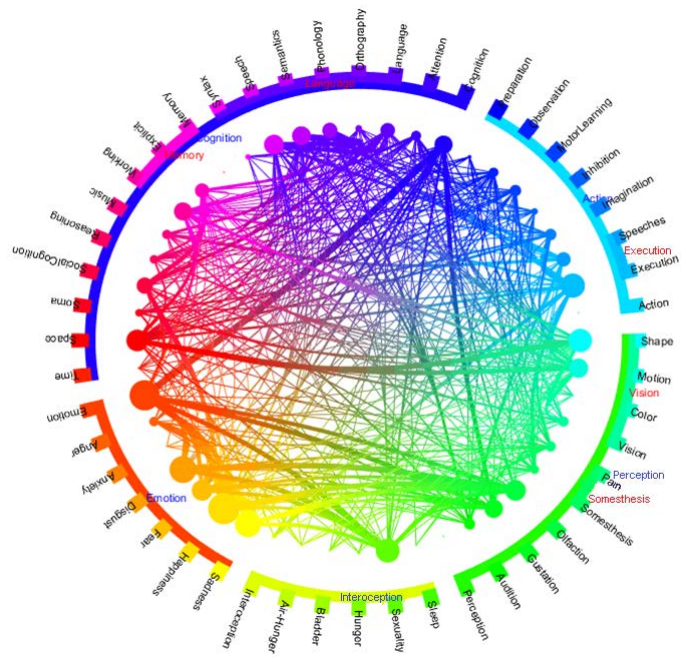


Fig. 3. Visualized analysis of 48 cognitive functions.

solving, planning, etc.). Cognitive science explores and studies the human thinking mechanism, especially the processing mechanism, by constructing cognitive models. It also provides a new architecture and technology for the design of intelligent systems.

In cognitive psychology, there are many cognitive frameworks such as ACT-R (Adaptive Control of Thought-Rational), SOAR (State, Operator And Result), ART (Adaptive Resonance Theory), synesthesia model, elementary perceiver and memorizer semantic network, human associative memory, GPS (General Problem Solver), PDP (Parallel Distributed Processing) and agent model. Among them, cognitive theory based on Bayesian probabilistic (see Appendix D) and perception, memory, and judgment mode [39] are very valuable cognitive architectures (see Appendix E). The computational theory based on the cognitive architecture can establish a cross-modal computing model of audio-visual media (see Appendix F).

Functional neuroimaging is an internal reflection of cognitive function, and it also is a technique for studying the cognition mechanism of the mind. Fig. 3 shows the results of our visual analysis of 48 cognitive functions in the neuroimaging database BrainMap (<http://www.brainmap.org/taxonomy>). BrainMap is a neuroimaging database of coordinate based functions and structures of the literature [40]-[42]. In general, the cognitive function of the human mind can be clearly found in the hierarchical structure.

Based on the visualization of cognitive function and the analysis, we proposed a framework for mind cognitive function (Fig. 4). Here, the cognitive process is mainly composed of perception pathway, motion controlled pathway, attentional controlled pathway, memory and emotion circuit, feeling and decision circuit, and judgment and control circuit etc.

D. Computable Theoretical Hypothesis of Brain and Mind

Fig. 5 is our visualized analysis of the correlation between 48 brain regions and 48 cognitive functions in the human brain LDP database according to the literature [35],[36]. In order to find the inherent law of neurocognitive, Fig. 6 is our simplification visualization of the correlation between 14 brain areas (includes 7 left brain areas and 7 right brain areas) and 5 cognitive functions in the literature [43],[44]. It can be noted that it is essentially fully connected between the cognitive function and neural connections. Frontal lobe and basal ganglion are

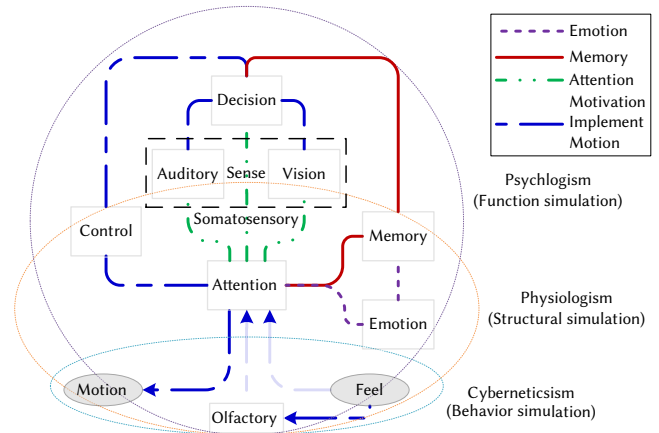


Fig. 4. Cognitive framework of mind.

the centers of neural processing, and the perception and emotional are the core of cognitive function. Cognitive function is closely related to the frontal temporal lobe, the thalamus and basal ganglion are closely related to the emotion, which is the projection center of information. These results are consistent with the basic theories of neuroscience and cognitive science.

The relationship between the structure of nervous system and the function of cognitive system is complex system of information control, and the relationship is unity of opposites. On the one hand, neural structure determines cognitive function; on the other hand, cognitive function also restricts neural structure. Both determine the system's intelligent behavior.

We can think that the structure of the brain nervous system and the function of the mental cognitive system constitute the "hardware" and "software" of the agent respectively, and the intelligent behavior of agents is generated by computational process. Fig. 7 shows the corresponding relationship among neural structure, cognitive function and intelligent behavior of the WMN in brain and mind. It is an isomorphic mapping between the brain structure of nervous system and mind function of cognitive system. Both of them reflected the different perspectives of the intelligent behavior.

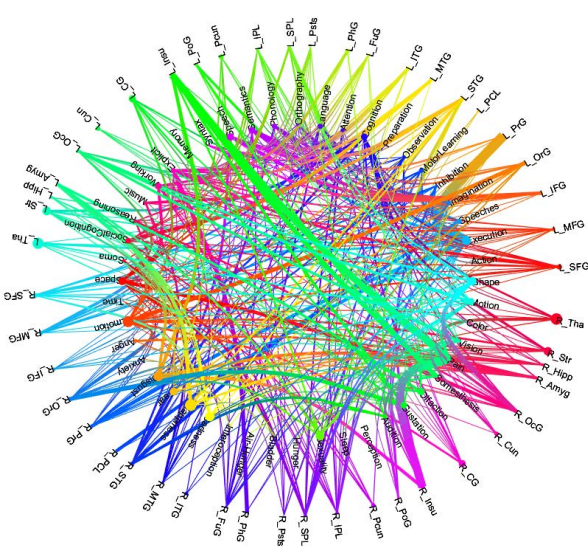


Fig. 5. Visualization and analysis between 48 brain regions and 48 cognitive functions.

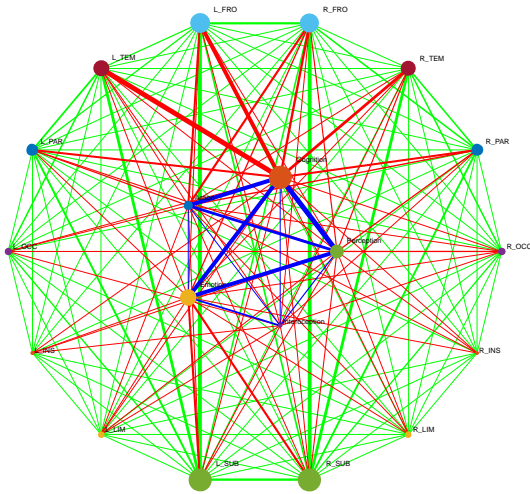


Fig. 6. Simplification visualization between 14 brain areas and 5 cognitive functions.

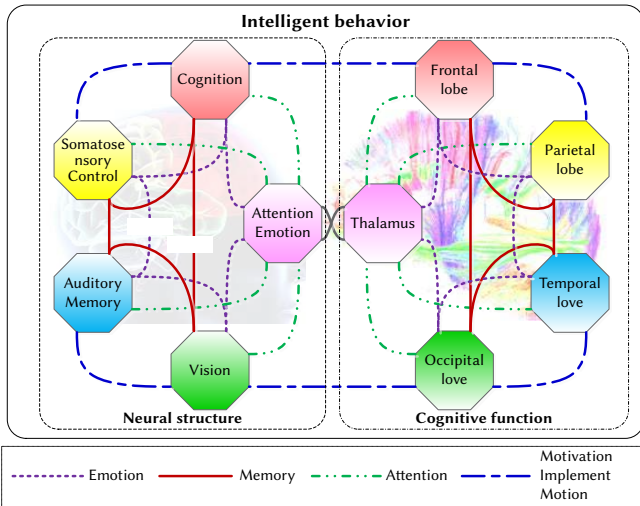


Fig. 7. The relationship among neural structure, cognitive function, and intelligent behavior.

According to the bionic principle and mechanistic, and based on the relationship between the structure, function, behavior and environment of natural intelligence, we proposes the following Cognitive Function and Information Hypothesis (CFI Hypothesis), Neural Structure and Control Hypothesis (NSC Hypothesis), and Complex System and Behavior Hypothesis (CSB Hypothesis) based on Function-Behavior-Structure(FBS) model [45] and SCI theory (system theory, cybernetics theory and information theory).

Hypothesis 1 (CFI Hypothesis). The mind function M comes from the organic combination of the modes of the cognitive system Ψ . As shown in Equation (1), cognitive function is produced by the orderly information processing by the system, which is an entropy reduction process (minimization of information entropy). Cognitive system can explore the function and processing mechanism of intelligence on a macroscopic level based on information theory.

$$M_i = \bigcup (\Psi_i | \sum F_k) \quad s.t. \quad argmin_p (-\sum p(f_k) \log p(f_k)) \quad (1)$$

Hypothesis 2 (NSC Hypothesis). The brain architecture B depends on the network structure of nervous system N . As shown in Equation (2), the dynamic optimization and control strategy of nervous system structure S is the result of multiple iterations and joint actions of target expectation Te , feedback information Fi and environmental interaction Ev . The nervous system can study the structure and control principle of intelligence on a microscopic level based on cybernetics.

$$B_i = \bigcap (N_i | \prod S_k) \quad s.t. \quad S_k \leftarrow Policy(S_{k-1}, Ev, Fi, Te) \quad (2)$$

Hypothesis 3 (CSB Hypothesis). Intelligent system is a complex nonlinear system with hierarchy. As shown in Equation (3), the intelligent behavior H is a process in which the module T processes the perceptual information layer by layer to realize the convex optimization computation C . Intelligent system can research the behavior and computational process of intelligence in mesoscopic level based on system theory.

$$H_i = C_i (C_{i-1} (\dots C_1 (T_1))) \quad s.t. \quad C_i \in Convex \text{ function} \quad (3)$$

According to the relationship between intelligent behavior, brain structure and mind function, Brain and Mind inspired Intelligence Hypothesis (BMI Hypothesis) can be proposed as follows.

Hypothesis 4 (BMI Hypothesis). Assuming that the nervous system N in brain B has structural S , the cognitive system Ψ in mind M has functional F , the natural system I has complex behavior H , and the computational model C has information processing process P , there is a homomorphic mapping Γ to realize brain and mind inspired intelligence (Equation 4).

$$\Gamma: \left\{ B \left(\bigcup (S|N) \right), M \left(\bigcap (F|\Psi) \right) \right\} \rightarrow I \left(\sum H \right) | C \left(\prod P \right) \quad s.t. \quad \begin{cases} \{S,N,B\} \cong \{F,\Psi,M\}, \{C,P\} \sim \{I,H\}, \\ \Psi \cap N = \emptyset, \Psi \subset M, N \subset B, I \subset C. \end{cases} \quad (4)$$

IV. KEY TECHNOLOGIES AND ENGINEERING APPLICATIONS OF BRAIN AND MIND INSPIRED COMPUTING

A. Semantic-Oriented MNCC Model

In view of the similarity neocortical structures and cooperation of cognitive function, we propose the following Brain and Mind Mechanism Hypothesis (BMM Hypothesis) and Target Classification and Recognition Hypothesis (TCR Hypothesis).

Hypothesis 5 (BMM Hypothesis). It can be assumed that the

information processing mechanism of the neocortex is universal in the brain areas. The audio-visual and other sensory procession can be modeled by uniform cortical function, and it can be applied to prediction, learning, reasoning and other general problem solver.

Hypothesis 6 (TCR Hypothesis). It can realize that the object semantics computing by BMC methods. That is, it needs to emulate the hierarchical processing, and attention mechanism of the nervous system in low-level. It also needs to imitate the framework of memory and emotion in middle-level, and simulate the function of probabilistic and causality reasoning based on the cognitive framework and integrated in high-level.

It is very necessary to establish unified hierarchical theory in mechanism of integrating behaviorism (or actionism), functionalism (or symbolism) and structuralism (or connectionism) [46]. Therefore, we constructed 4 layers Multimedia Neural Cognitive Computing (MNCC) model for semantics-oriented computing of BMI. Each of the layers is described as follows:

- Layer 0 (Hybrid computation layer based on mathematical model of endocrine, immune and neurochemical).

The hybrid computation layer emulates biochemical brain intelligent by Artificial Endocrine System (AES), Artificial Immune System (AIS) and NeuroChemical System (NCS) such as necrohormones, neurotransmitter and neuromodulator (or neuropeptide).

- Layer 1 (Perceptual computation layer based on control models).

The perceptual computation layer bionics realizes the cognitive function of perception and attention, which formed by the neural structure of thalamus, primary cortex of temporal lobe, parietal lobe and occipital lobe. The perceptual computation layer imitates motor brain intelligent of perceived behavioral control on archicortex.

- Layer 2 (Neural computation layer based on structural models).

The neural computation layer bionics realizes the cognitive function of memory, emotion and sensation, which formed by the neural structure of thalamus, secondary cortex of temporal lobe, parietal lobe and occipital lobe. The neural computation layer imitates emotional brain attention circuit, emotional circuit and memory circuit of the limbic system on paleocortex. The models of this layer had incremental learning based on emotion computing, reinforcement learning based on memory, deep learning such as SNN, DBN and CNN et. al.

- Layer 3 (Cognitive computation layer based on functional models).

The cognitive computation layer bionics realizes the cognitive function of perception, inference, prediction and judgment, which formed by the neural structure of frontal lobe, association cortex of temporal lobe, parietal lobe and occipital lobe. The cognitive computation layer simulates rational brain of hierarchical ensemble learning, subjective Bayesian cognitive learning, language, and thinking control in neocortex. The models of this layer had HMM, LDA, PGM et. al.

Semantic-oriented computing needs to research and discover the cortex structure of the nervous system, the network structure of the white matter, and the cognition function of the mind, such as hierarchical processing, incremental memory, emotional reinforcement, probability ensemble and so on.

As Fig. 8 shows, a semantic-oriented MNCC model based on the neural structure and cognitive framework were proposed. MNCC model is designed based on the characteristics of neural cognitive information processing such as information transmission and feedback, hierarchical, distributed and parallel processing. It extracts semantic information from the representation media by multiple

steps such as a region of interest (ROI) extraction, saliency target detection, object-oriented incremental recognition, multi-scale target reinforcement, hierarchical ensemble process and other steps.

B. Hierarchical CCNC Framework

In view of the hierarchical of the natural media such as audio and video, high-level features can be achieved through the combination of low-level features. There are hierarchical structures in the language text such as words, sentences, paragraphs, and documents. There are hierarchical structures in a speech sound, for instance, sampling, phonemes, syllables, and words. Similarly, there are hierarchical structures in the natural images, for example, pixels, edges, shapes, textures, objects and scenes. From the related research of cognitive science and neuroscience, the information processing of cognitive function and neural structure also has a similar hierarchical structure. Considering the hierarchy of neural cognitive for semantic computing, Fig. 9 is our further improved hierarchical CCNC framework based on MNCC.

The hierarchical CCNC framework is designed based on the Brain and Mind inspired Computing Hypothesis (BMC Hypothesis) as follows.

Hypothesis 7 (BMC Hypothesis): The computing system can be layered in hybrid computation, perceptual computation, neural computation, and cognitive computation to realize general artificial intelligence. That is, it can emulate low-layer perception computing process based on saliency mechanism and swarm intelligence. It can imitate the middle-layer of hierarchical feature computing process based on deep learning, reinforcement learning, and incremental learning. It also can simulate high-layer hierarchical decision process based on probability reasoning, causality reasoning, and ensemble learning.

The goal of CCNC mainly solves the problems of multimodal semantic and cross-modal computing. Based on the 4 layers of the MNCC model, the CCNC framework extends it into 7 sub-layers and 1 hybrid layer, which can realize the semantic computing function. Each layer is described as follows:

L₀ Hybrid computation layer

This layer is designed and implements the dynamic I/O regulation according to the prior rules and inhibition and excitation mechanism of AES and AIS. It can also inhibit or excite to other layers by NCS.

L₁ Perceptual computation layer

L_{1.1} This sub-layer realizes pre-processing of perceptual information.

L_{1.2} This sub-layer imitates attention mechanism of the thalamus-cortical circuit and extracts the saliency features from the media target based on sparse representation.

L₂ Neural computation layer

L_{2.1} This sub-layer emulates the hierarchical structure of cortical columns, and constructs the semantic classifier based on deep learning.

L_{2.2} This sub-layer emulates the emotional reward and punishment mechanism of the limbic system, and realizes the function of the semantic reinforcement learning.

L_{2.3} This sub-layer emulates the memory mechanism of the cortex-hippocampus system and realizes the function of the incremental semantic learning.

L₃ Cognitive computation layer

L_{3.1} This sub-layer simulate the theory of mono-modal cortical column and Bayes subjective probability to realize semantic cognition computing.

L_{3.2} This sub-layer simulate the information integration multi-modal cortical column to realize semantic ensemble learning of multiple classifiers.

TABLE I. THE SYMBOL OF CCNC FRAMEWORK AND ITS IMPLICATION

Symbol	Types	Implication
MM	Multidimensional matrix set	Media set(include image, audio, text, and video)
Ma	Tensor	Media data
SC	Algorithm	Saliency computation
Sa	Sparse tensor	Temporal saliency feature (sparse representation)
Sv	Sparse tensor	Spatial saliency feature (sparse representation)
DL	Algorithm	Deep learning algorithm
IL	Algorithm	Incremental learning algorithm
RL	Algorithm	Reinforcement learning algorithm
EL	Algorithm	Ensemble learning algorithm
CC	Algorithm	Cognitive computing algorithm
Cp	Set	Target semantics
Ct	Vector	Features of temporal perception (probability topic)
Cs	Vector	Features of spatial perception (probability topic)
Fa	Sparse matrix	Features of temporal senses (probability distribution)
Fv	Sparse matrix	Features of spatial senses (probability distribution)
Ma	Sparse matrix	Time increment of DNN
Mv	Sparse matrix	Space increment of DNN
Mt	Vector	Time increment of cognitive topic
Ms	Vector	Space increment of cognitive topic
Mp	Vector	Feedback information of incremental learning
Mn	Vector	Attention increment of saliency computation
Es	Parameter	Reinforcement feedback of saliency computation
Ei	Parameter	Incremental feedback of memory
EH	Set	Endocrine molecules which effect on input and output
SS	Set	Semantic state of chemical solution
TS	Mapping	Reaction rule

The hierarchical CCNC framework can be described by following the 7-tuple $\langle SC, EL, IL, RL, DL, CC, EH \rangle$. It is mapping processing that the CCNC framework training and recognition, which can be described following Equation (5).

$$\begin{aligned} CCNC: (MM|MNCC) \rightarrow Cp \\ s.t. \quad CCNC = \langle SC, EL, IL, RL, DL, CC, EH \rangle \end{aligned} \quad (5)$$

The symbols and illustration of the hierarchical CCNC framework in Fig. 9 and Formula 5 are shown in Table I.

The formal semantics of CCNC framework based on CHemical Abstract Machine (CHAM) presented in Appendix G. The algorithms description for semantic learning and recognition of hierarchical CCNC framework presented in Appendix H.

C. System Applications of CNCC Architecture

A wide range of applications of semantic-oriented MNCC model and hierarchical CCNC framework would be identified such as unmanned autonomous system and search engines of cross-media intelligent [53]. It would have profound significance for the exploration and the realization of the BMC. With the development of software defining satellite, on-board software urgently needs high productivity computing to solve the problem of remote sensing intelligent information extraction.

We established a Cross-modal Neural Cognitive Computing (CNCC) architecture based on MNCC model and CCNC framework. CNCC can provide high productivity intelligent algorithms and toolkits for remote sensing information extraction. As Fig. 10 shows, CNCC architecture had been applied to the algorithms of scene classification, target detection and target recognition of high-resolution remote sensing images [51].

V. DISCUSSION

In order to verify the application performance based on BMC in remote sensing information extraction. Table II lists the experimental results of our previous research on scene classification, target detection and target recognition based on the CNCC architecture.

TABLE II. THE EXPERIMENTAL RESULTS SEMATIC RECOGNITION BASED ON CNCC ARCHITECTURE.

Semantic recognition	Datasets	Model / Method	AP (%)	OA (%)	PD (%)	FAR (%)	MDR (%)
Scene classification [47]	HRSS	MNCC/ SC-MNCC	84.73				
Scene classification [47]	UCMLU	MNCC/ SC-MNCC	88.26				
Sonar target classification [48]	SITC	CCNC/ SABP	91.11				
Target detection [49]	HRSHTD	MNCC/ SLS-CNN			95.00	8.00	5.00
Ship detection [50]	SAR	CCNC/ SD-SNN			91.63	9.48	11.02
Tank recognition [51]	MSTAR	CCNC/ TCR-EL-DHMM	99.90	99.88			
Target recognition [51]	HSTCR	CCNC/ TCR-IREL-OOMS	97.00	96.93			
Hyperspectral image classification [52]	IP	CCNC/ SABP	99.31	98.21			

The Average Precision (AP) of scene classification algorithm based on MNCC model is up to 84.73% on High-resolution Satellite Scene (HRSS) dataset of Wuhan University and reaches 88.26% on University of California Merced Land Use (UCMLU) dataset. For target detection algorithm based on MNCC model on High-resolution Remote Sensing Harbor Target Detection (HRSHTD) dataset, the average Probability of Detection (PD) is 91.63%, False Alarm Rate (FAR) is 8.37%, and Missed Detection Rate (MDR) is 9.35%.

The experimental results show that AP and Overall Accuracy (OA) of target classification algorithm based on CCNC framework are 96.93% and 97.00% on High-resolution Ship Target Classification and Recognition (HSTCR) dataset, respectively [51]. It also reaches 99.90% and 99.88% on Moving and Stationary Target Acquisition and

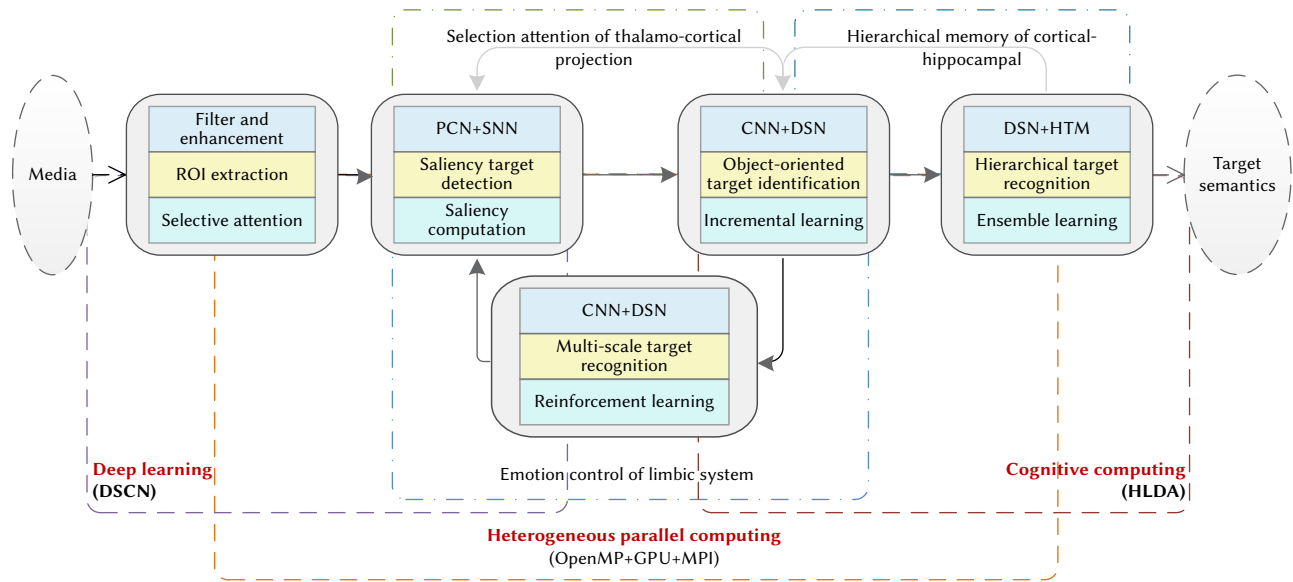


Fig. 8. The semantic-oriented MNCC model.

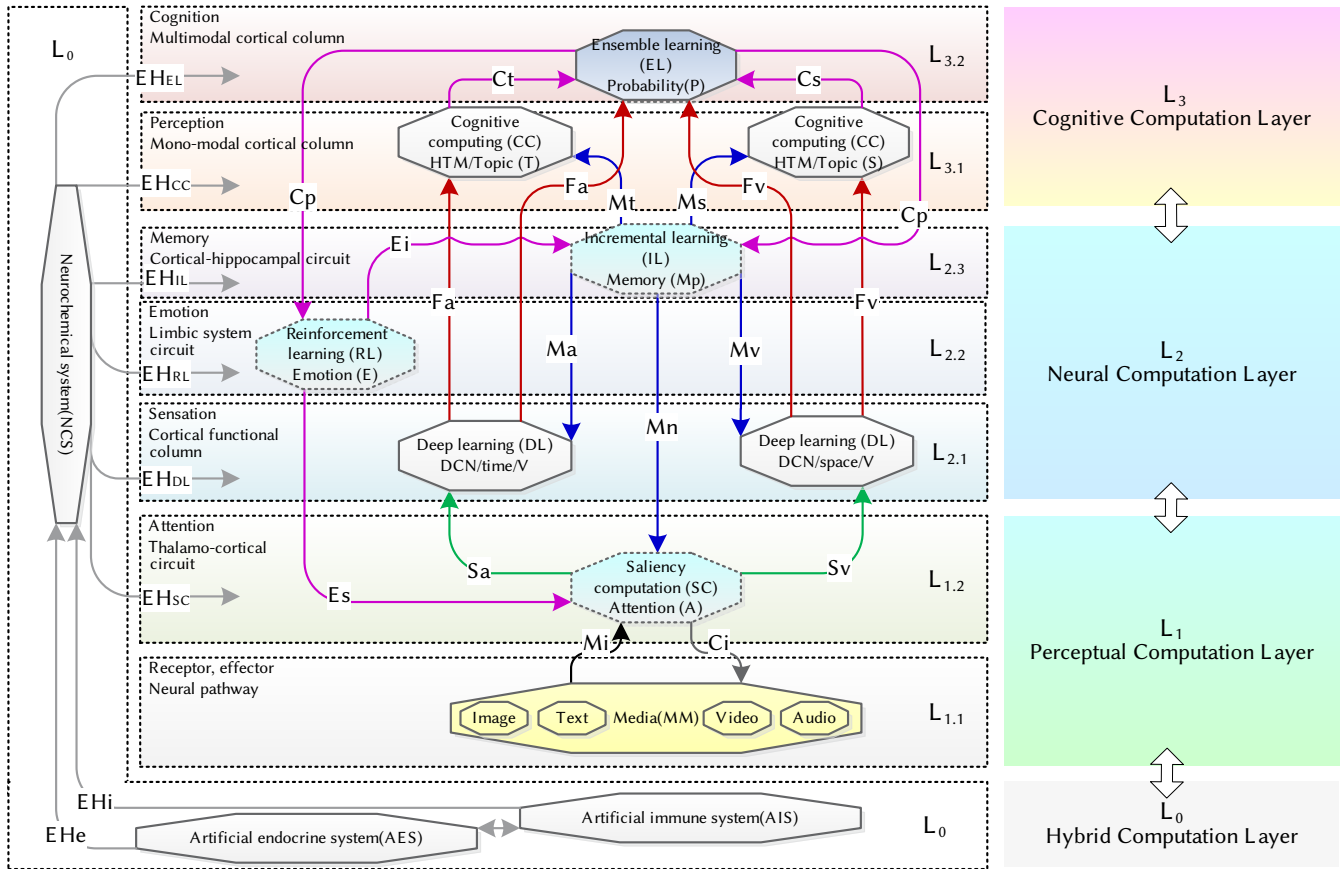


Fig. 9. The hierarchical CCNC framework.

Recognition (MSTAR) dataset, respectively. It shows that the CNCC architecture can address the problem of semantic learning on remote sensing image, which is complex ground objects.

This research shows that the semantic oriented MNCC model and the hierarchical CCNC framework designed by us based on brain mechanism and mind architecture can effectively improve the semantic processing performance of multimedia and cross-modal information, such as target detection, target classification and target recognition.

VI. CONCLUSION

To address the problems of scientific theory, common technology and engineering application of multimedia and multimodal information computing, we are focused on the theoretical model, algorithm framework and system architecture of BMI based on the structure mechanism simulation of nervous system, the function architecture emulation of cognitive system and the complex behavior imitation of natural system.

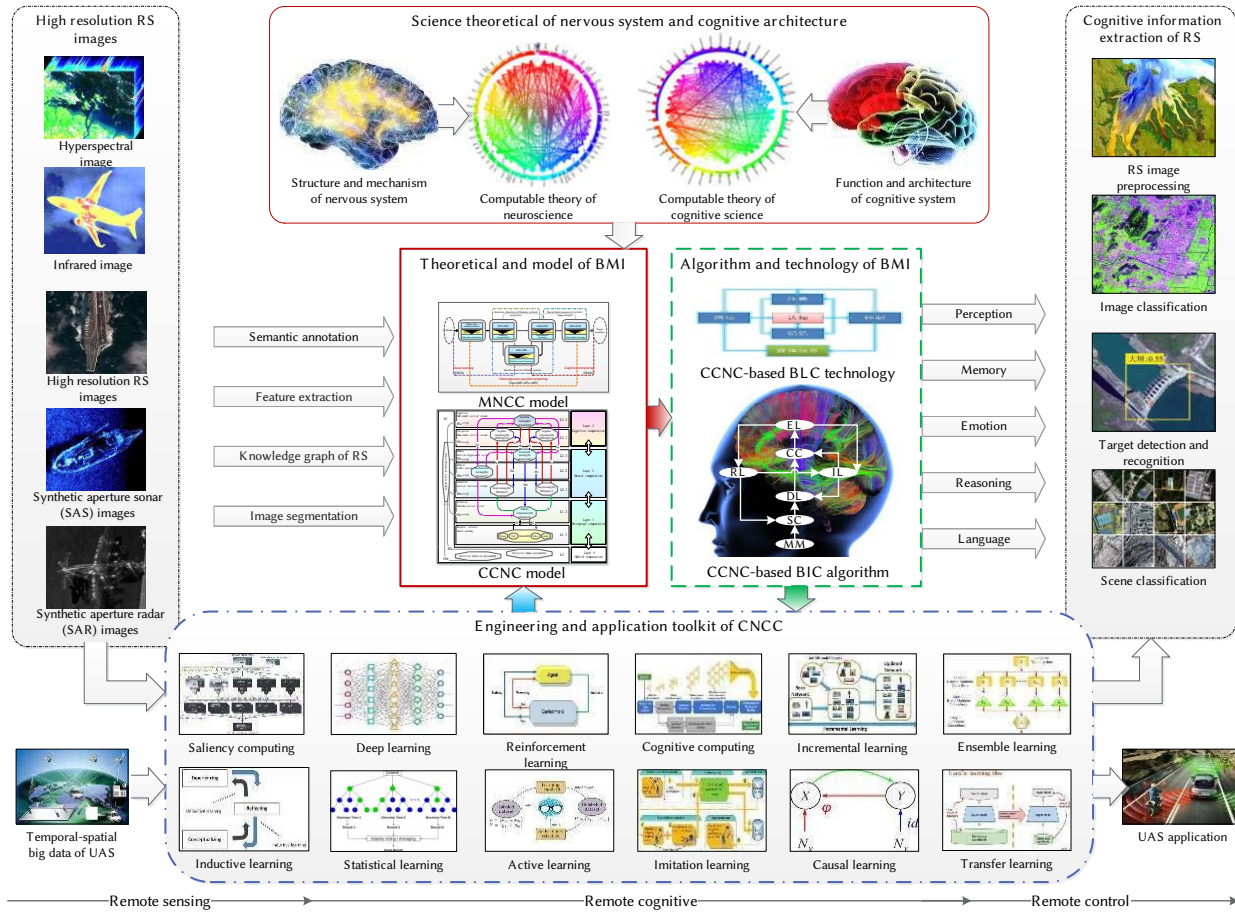


Fig. 10. The system application of the CNCC architecture for information extraction of remote sensing and target recognition of unmanned autonomous system.

Based on information theory, system theory, cybernetics and bionics, we define the concepts of BMC and propose the assumptions of BMI. Aiming at scientific problems of BMI modeling, the cortical models and nervous system structure in human brain WMN had been analyzed; the hierarchy characteristic of architecture and cognitive system function in mind had been explored. The relationship between nervous system and cognitive framework for BMI had also summarized. Then hierarchical CCNC framework is proposed based on the MNCC model, and the rationality of the hierarchical CCNC framework is formally analyzed based on CHAM. The semantic learning and recognition algorithm of our models are given. Our research on remote sensing intelligent information extraction and cross-media information retrieval shows that the scene classification, target detection, target classification and target recognition based on BMC algorithm have very high performance.

The BMI theory proposed can be widely used in high-resolution earth observation system and cross-media search engine and other applications. Looking to the future, CNCC architecture will be applied to more cross-modal intelligence information perception of unmanned autonomous systems and platforms, such as Unmanned Ground Vehicle (UGV), Unmanned Aerial Vehicle (UAV), Unmanned Surface Vehicle (USV), Unmanned Underwater Vehicle (UUV), Software Defined Satellite (SDS), intelligent robot and other unmanned autonomous equipment. The next step is to improve the BMI theoretical system, overcome the key technologies of BMC, and realize the state-of-the-art application of complex systems based on CNCC architecture.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (No. 62176087) , Shenzhen Special Foundation of Central Government to Guide Local Science & Technology Development (No. 2021Szvup032), Postgraduate Education Reform and Quality Province Improvement Project of Henan Province (No. YJS2022JC33).

APPENDIX

A. Cortical Model

There are three types of nervous system models: description model of nervous system, neural mechanism model, and interpretation model of neural function. The description model quantitatively describes nervous system based on the experimental data. The mechanism model emulates nervous system how to run. The interpretation model explores the basic principles of the nervous system, and the construction of the nervous system why so run. Typical nervous system models include neuron model, synaptic model, cortical model and structural model of the nervous system.

According to the evolutionary hypothesis of the triune brain [54], Paul MacLean divides the model of human brain structure and function into 3 specific regions: archicortex, paleocortex and neocortex. The archicortex originates from motor brain (reptilian), which cortical structure is not very obvious. The paleocortex of the emotional brain (paleomammalian) lies in limbic system consists of 3 layers of neurons. The neocortex consists of 6 layers of neurons, which accounted for 90% of the area of the rational brain (neomammalian).

The triune brain hypothesis is a controversial and extremely simplified model [54], [55]. Generally, the neocortex can be divided into primary areas, secondary areas, association areas in function.

The layer structure of 3 types cortical areas [55], [56] as showed in Fig. 11. The research shows that 6 layers of the neocortex have different functions. For example, the L4 layer receives inputs information. The L2 and L3 layers make up a local circuit for information processing. The L1 layers achieve intersection and inhibition projection information of internal neurons, and information output from the L5 and L6 layers.

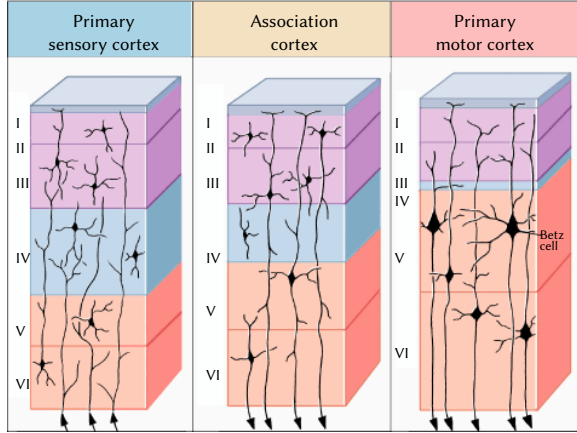


Fig. 11. Cortical structure.

Most studies suggest that neocortex has the similar structure in vision area, audition area, and association area. Cortical columns are a basic unit for information processing in neocortex. Cortical columns have the phenomenon of hierarchical processing and the mechanism of lateral inhibition of each other. Micro-columns consist of local circuits in neocortex. Physical stimuli are perceived and encoding to generate neural spiking coding by visual-auditory sensory neurons. The micro-column is feature detection, and macro-column or super-column makes up of micro-columns to process special information and generates some cognitive functions. The spiking probability is propagating among micro-columns. Micro-columns collect information from lower neighbor micro-columns and disseminate information from upper neighbor micro-columns [43]. At the same time, it also receives feedback information from LDP, and prediction information from an upper neighbor.

1. Temporal-Spatial Structure of Micro-Column Node

For simplicity in the model design, we firstly merge micro-column with 6 nodes (Fig. 12(a)) to micro-column with 3 nodes (Fig. 12(b)). The middle layer (L4) receives the input information. The lower layers (L5 and L6) send output information, and the upper layers (L1, L2, L3) process information. In fact, cortex information processing has the spatial-temporal property. So we further simplify the model structure with 2 nodes. It is noted that this simplification does not lose the advantage of bionics. For instance, the node is double structure in HTM, RBM, SVM and so on. As shown in Fig. 12(c), S mimicking functions from L1 to L4, and simulates memory and spatial patterns process [43]. T mimicking functions of L5 and L6, and simulates memory and temporal patterns process. Both nodes S and T memories belief which comes from owner and other nodes.

2. Hierarchical Network Architecture of Super-Column

According to neurocognitive system hierarchical architecture and temporal-spatial locality, super-column architecture also uses hierarchical, multi-level, and bidirectional mapping structure. Super-column composed by a micro-column with principles of “the same layer collaborative” and “hierarchical processing” (Fig. 13) [43].

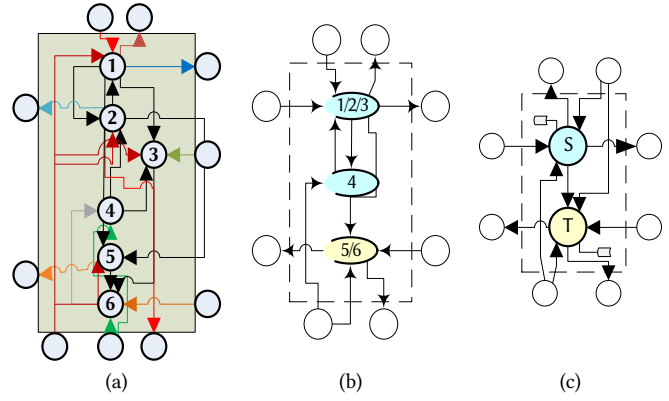


Fig. 12. The hierarchical structure model of the micro-column. (a) Micro-column structure with 6 nodes. (b) Micro-column structure with 3 nodes. (c) Micro-column structure with 2 nodes.

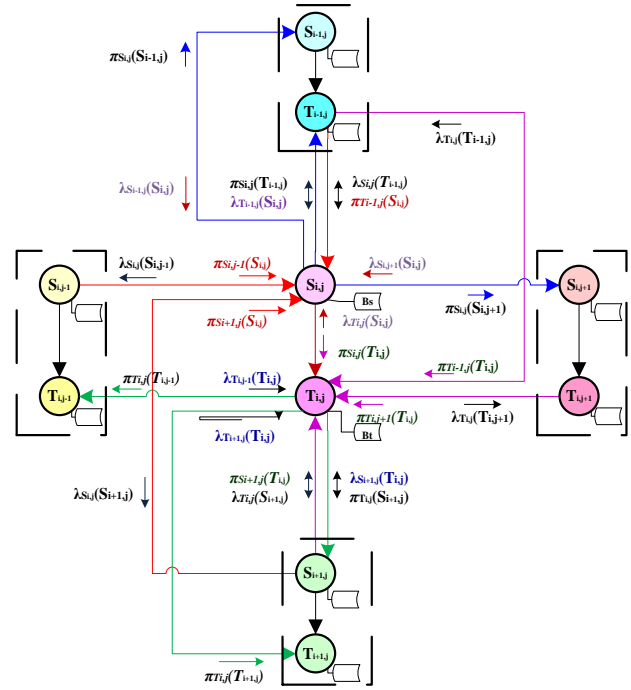


Fig. 13. The structure model of the super-column.

B. Long Distance Pathways and Neural Circuits

According to the whole brain LDP database, the frontal lobe has the core nodes, and the thalamo-cortical projection system is the key connection in network structure of the human’s brain. Both human visual system and human auditory system have a dual stream model: dorsal and ventral pathway, as shown in Fig. 14. “What” is happening in the dorsal pathway, and “where” is happening in the ventral pathway.

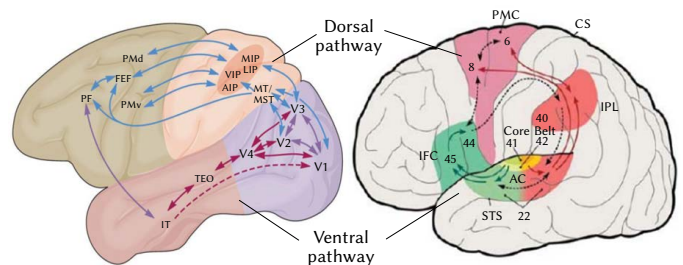


Fig. 14. The dorsal and ventral pathway in human visual system (left) and human auditory system (right).

The high accuracy human brain LDP database based on the experimental data taken from documents was constructed. Fig. 15 shows the pathways model in human visual system and human auditory system [35]-[38].

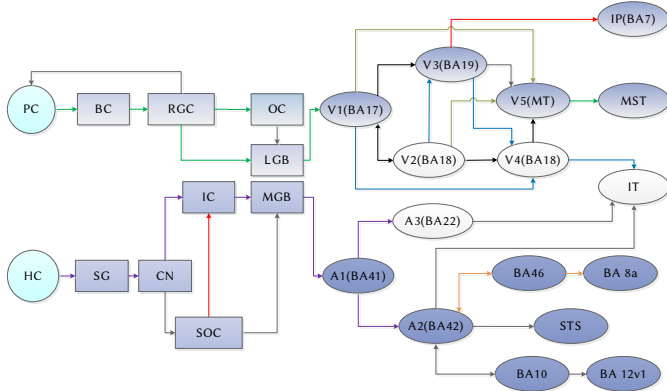


Fig. 15. The processing pathway model of human visual system (up) and human auditory system (down).

C. Hierarchical Temporal Memory

Hierarchical Temporal Memory (HTM) model [34] is a kind of neocortex structure and function by Jeff Hawkins and Dileep et al. It adopted Bayesian Belief Propagation (PBP) theory to explain the neocortex process of recognition and reconstruction (Fig. 16). In order to further simulate the structure of the neocortex, the concept of the cognitive domain of sparse distributed representation and other neuroscience concepts, such as the dendrites, synapses, and so on, is introduced. It proposed Cortical Learning Algorithm (CLA) in HTM, and its fundamental idea is hierarchy structure, and invariant representations of spatial patterns and temporal patterns and sequence memory.

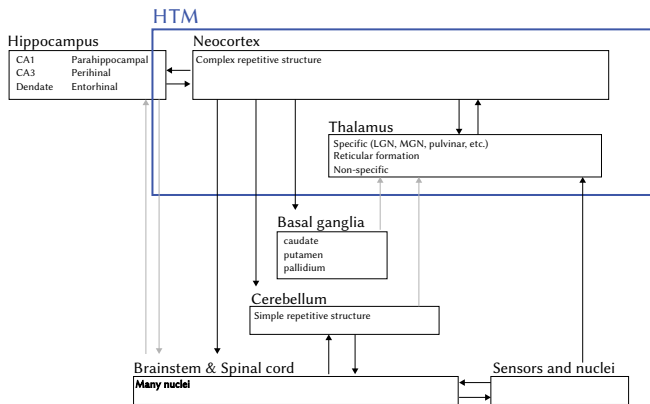


Fig. 16. HTM model.

D. Cognitive Theory Based on Bayesian Probabilistic

Since the Bayesian proposed the probabilistic theory in 1963, the probability reasoning and decision-making of the uncertainty information had become an important content of the researches on the objective probability and cognitive processing. Bayesian rule describes the likelihood between the priori probability (marginal probability) $P(x)$ and the posterior probability (conditional probability) of the historical information $P(x_i|x_j)$. The Bayesian rule provides a method for modifying and reasoning about the probability distribution of the subjective judgment $P(x_j)$ for observed phenomena. If x_i, x_j is condition independent, the sum-product rule can be derived by Bayesian inference as follows Equation (6).

$$\begin{cases} \text{Sum rule: } P(x_j) = \sum_{i=1,2,\dots,n,i \neq j} P(x_j, x_i) \\ \text{Product rule: } P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i | \pi(x_i)) \\ \text{s. t. } P(x_i | x_j) = \frac{P(x_j | x_i) P(x_i)}{P(x_j)}, P(x_i x_j) = P(x_i) P(x_j) \end{cases} \quad (6)$$

It is the mainstream method of machine learning and reasoning depending on the uncertainty representation of the probability, the Bayesian rule, and the extension model. Cognitive researchers use Bayesian brain model [57] to simulate the cognitive process and model of mind. It is investigated cognitive processing law of subjective probability estimation by a probability model. Bayesian brain theory believes that the brain is a predictive machine, and cognition is the process of probability calculation.

E. Perception, Memory and Judgment Model

Cognitive science researchers think that between the cognitive of the human mind and the computer information is similar in processing process. They are establishing cognitive computing theory according to computers to simulate human cognitive processes. It is research and analyses the processes and principles of human cognition, discover the main stages and pathways of cognitive processes, and establish the relationship between cognitive processes and computing workflow.

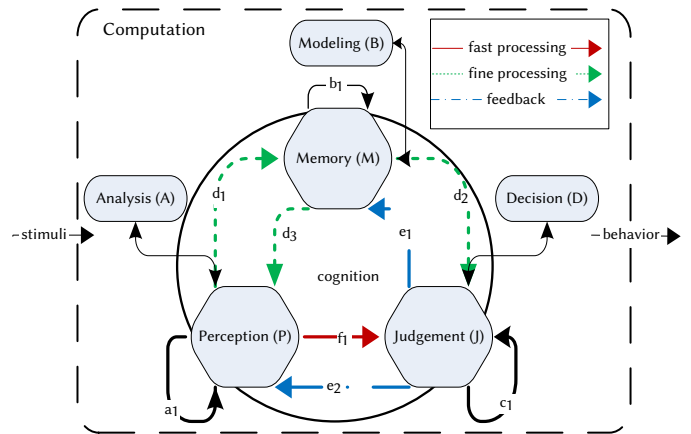


Fig. 17. PMJ model.

Fig. 17 is a cognitive computing model, which is constructed based on Perception, Memory and Judgment (PMJ) model from literature [39], [58]. Cognitive processing mainly consists of 3 stages which include perception (a_i), memory (b_i) and judgment (c_i) in PRJ model. There are 3 pathways, summarized as the fast processing pathway (f_1), the fine processing pathway (d_1, d_2 and d_3), and the feedback processing pathway (e_1 and e_2). The perception, memory and judgment of the cognitive process are respectively corresponding with time dependent mapping that the analysis of the computational process (A), modeling (B) and decisions (D) as follows Equation (7).

$$\begin{aligned} pmj: & \langle P_t, M_t, J_t \rangle \rightarrow \langle A_t, B_t, D_t \rangle \\ \text{s. t. } & \begin{cases} p: \langle P_t, M_t, J_t, P_{t-1} \rangle \rightarrow A_t \\ m: \langle P_t, M_t, J_t, M_{t-1} \rangle \rightarrow B_t \\ j: \langle P_t, M_t, J_t, J_{t-1} \rangle \rightarrow D_t \end{cases} \end{aligned} \quad (7)$$

F. Cognitive Architecture for Media Computing

As Fig. 18 shows that the cognitive framework for brain-inspired processing of audio-visual can be divided into four steps [44]. That is, computation and simulation of cortical columns belief, computation and simulation of control information of thalamus for attention, computation and simulation of control information of limbic system for emotion, and computation and simulation of control information of spatio-temporal semantic caching of hippocampus for memory.

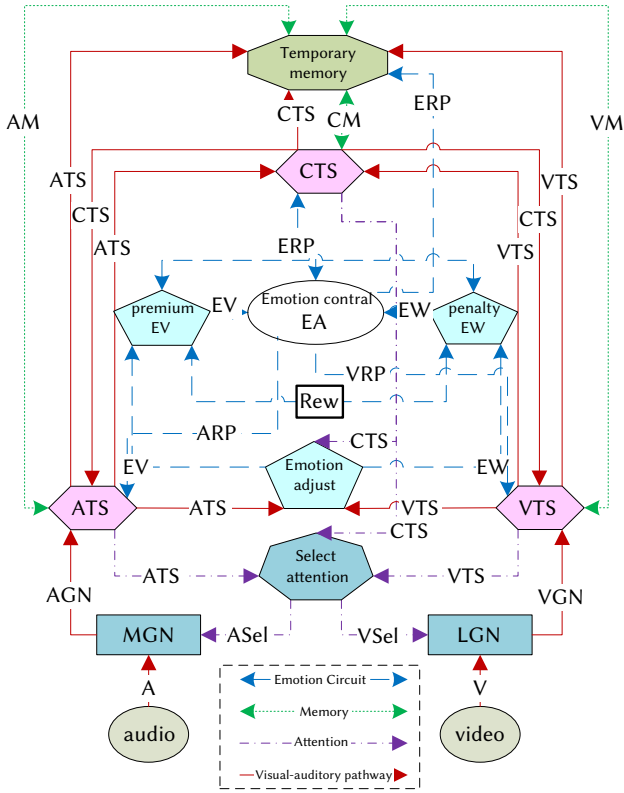


Fig. 18. Cognitive architecture of media computing.

The whole strategy of the media information processing process is training and reinforcement layer by layer. It includes 2 steps pre-learning algorithm in waking (PLAW) and precisely adjust algorithm in sleeping (PAAS) as follows:

1. PLAW mimics cognitive function is controlled by emotion and memory under the waking state. It is unsupervised training and using bottom-up and to pre-process input information of temporal media A and spatial media V step by step, and generate a set of MNCC initial parameters.
2. PAAS mimics cognitive function of sleeping state when thalamus closed the input information of temporal media A and spatial media V. It is supervised training, and top-down adjusts and optimizes internal parameters with hierarchical reinforcement learning strategies under the memory and emotional control.

The reward function of the limbic system is designed by the “principle of lowest energy” and “maximizing benefit” of the system. That is, rewarding successes and punishing failure.

G. Formal Semantics of CCNC Framework Based on Chemical Abstract Machine

The mind and brain is the physical and chemical reaction of biology. In order to analyze rationality of CCNC architecture, the Chemical Abstract Machine (CHAM) [59],[60] was employed. CHAM is a kind of description language architecture for parallel and dynamic software architecture analysis and testing. CHAM describes intelligent system architecture with molecules EH (e.g., hormones, neurotransmitters, and receptors), solution SS (e.g., state, semantic) and rules TS (e.g., knowledge, association, mapping). The CHAM molecular EH denoted factors of the chemical systems such as hormones, receptors, and transmitters, which affect the function of the physical system in the nervous system and cognitive architecture.

$$EH = EH_{SC}, EH_{DL}, EH_{RL}, EH_{IL}, EH_{CC}, EH_{EL}$$

The connecting elements C, processing elements TS (such as knowledge, rule, association, and mapping) and data elements D was defined as follows:

$$M::=TS|C\Diamond EH|EH\Diamond C|EH\Diamond EH$$

$$C::=i(D)|o(D)|g(EH)|d(EH)$$

$$TS::=SC|IL|EL|RL|DL|CC$$

$$D::=Mi|Sa|Sv|Fa|Fv|Ma|Mv|Ms|Mt|Mp|Ei|Es|Cs|Ct|Cp|EH$$

where $i(\cdot)$ denoted the input, $o(\cdot)$ denoted the output, $g(\cdot)$ denoted the effects on the system input of the generation of hormones and transmitters, $d(\cdot)$ denoted the effects on the system output of the receptors receiving hormone and transmitter. The initial solution SS was defined as follows:

$$SS = SS_{SC} // SS_{DL} // SS_{CC} // SS_{EL} // SS_{RL} // SS_{IL}$$

where sub-solution is denoted as follows:

$$SS_{SC} = \{i(Mi) \Diamond i(Mn) \Diamond i(Es) \Diamond g(EH_{SC}) \Diamond SC \Diamond o(Sa) \Diamond o(Sv) \Diamond d(EH_{SC})\}$$

$$SS_{DL} = \{i(Sa) \Diamond i(Ma) \Diamond g(EH_{DL}) \Diamond DL \Diamond o(Fa) // i(Sv) \Diamond i(Mv) \Diamond DL \Diamond o(Fv) \Diamond d(EH_{DL})\}$$

$$SS_{RL} = \{i(Cp) \Diamond g(EH_{RL}) \Diamond RL \Diamond o(Ei) \Diamond o(Es) \Diamond d(EH_{RL})\}$$

$$SS_{IL} = \{i(Ei) \Diamond i(Cp) \Diamond g(EH_{IL}) \Diamond IL \Diamond o(Mp) \Diamond o(Ma) \Diamond o(Mv) \Diamond o(Mt) \Diamond o(Ms) \Diamond d(EH_{IL})\}$$

$$SS_{CC} = \{i(Fa) \Diamond i(Mt) \Diamond g(EH_{CC}) \Diamond CC \Diamond o(Ct) \Diamond d(EH_{CC}) // i(Fv) \Diamond i(Ms) \Diamond g(EH_{CC}) \Diamond CC \Diamond o(Cs) \Diamond d(EH_{CC})\}$$

$$SS_{EL} = \{i(Ct) \Diamond i(Cs) \Diamond i(Fa) \Diamond i(Fv) \Diamond g(EH_{EL}) \Diamond EL \Diamond o(Cp) \Diamond d(EH_{EL})\}$$

The intermediate solution SM after the reaction was defined as follows:

$$SM = SM_{SC} // SM_{DL} // SM_{CC} // SM_{EL} // SM_{RL} // SM_{IL}$$

where the sub-solution is denoted as follows:

$$SM_{SC} = \{SC \Diamond i(Mi) \Diamond i(Mn) \Diamond i(Es) \Diamond g(EH_{SC}) \Diamond o(Sa) \Diamond o(Sv) \Diamond d(EH_{SC})\}$$

$$SM_{DL} = \{DL \Diamond i(Sa) \Diamond i(Ma) \Diamond g(EH_{DL}) \Diamond o(Fa) // DL \Diamond i(Sv) \Diamond i(Mv) \Diamond o(Fv) \Diamond d(EH_{DL})\}$$

$$SM_{RL} = \{RL \Diamond i(Cp) \Diamond g(EH_{RL}) \Diamond o(Ei) \Diamond o(Es) \Diamond d(EH_{RL})\}$$

$$SM_{IL} = \{IL \Diamond i(Ei) \Diamond i(Cp) \Diamond g(EH_{IL}) \Diamond o(Mp) \Diamond o(Ma) \Diamond o(Mv) \Diamond o(Mt) \Diamond o(Ms) \Diamond d(EH_{IL})\}$$

$$SM_{CC} = \{CC \Diamond i(Fa) \Diamond i(Mt) \Diamond g(EH_{CC}) \Diamond o(Ct) \Diamond d(EH_{CC}) // CC \Diamond i(Fv) \Diamond i(Ms) \Diamond g(EH_{CC}) \Diamond o(Cs) \Diamond d(EH_{CC})\}$$

$$SM_{EL} = \{EL \Diamond i(Ct) \Diamond i(Cs) \Diamond i(Fa) \Diamond i(Fv) \Diamond g(EH_{EL}) \Diamond o(Cp) \Diamond d(EH_{EL})\}$$

The 6 important basic rules for the solution reaction (state transition) were defined as follows:

$$TS_{SC} = i(Mi) \Diamond i(Mn) \Diamond i(Es) \Diamond g(EH_{SC}) \Diamond SC, o(Sa) \Diamond o(Sv) \Diamond d(EH_{SC}) \Diamond SC \rightarrow SC \Diamond i(Mi) \Diamond i(Mn) \Diamond i(Es) \Diamond g(EH_{SC}), SC \Diamond o(Sa) \Diamond o(Sv) \Diamond d(EH_{SC})$$

$$TS_{DL} = i(Sa) \Diamond i(Ma) \Diamond g(EH_{DL}) \Diamond DL, o(Fa) \Diamond d(EH_{DL}) \Diamond DL, i(Sv) \Diamond i(Mv) \Diamond g(EH_{DL}) \Diamond DL, o(Fv) \Diamond d(EH_{DL}) \Diamond DL \rightarrow DL \Diamond i(Sa) \Diamond i(Ma) \Diamond g(EH_{DL}), DL \Diamond o(Fa) \Diamond d(EH_{DL}), DL \Diamond i(Sv) \Diamond i(Mv) \Diamond g(EH_{DL}), DL \Diamond o(Fv) \Diamond d(EH_{DL})$$

$$TS_{RL} = i(Cp) \Diamond g(EH_{RL}) \Diamond RL, o(Ei) \Diamond o(Es) \Diamond d(EH_{RL}) \Diamond RL \rightarrow RL \Diamond i(Cp) \Diamond g(EH_{RL}), RL \Diamond o(Ei) \Diamond o(Es) \Diamond d(EH_{RL})$$

$TS_{IL} = i(Ei) \diamond i(Cp) \diamond g(EH_{IL}) \diamond IL, o(Mp) \diamond o(Ma) \diamond o(Mv) \diamond o(Mt) \diamond o(Ms) \diamond d(EH_{IL}) \diamond IL \rightarrow IL \diamond i(Ei) \diamond i(Cp) \diamond g(EH_{IL}), IL \diamond o(Mp) \diamond o(Ma) \diamond o(Mv) \diamond o(Mt) \diamond o(Ms) \diamond d(EH_{IL})$

$TS_{CC} = i(Fa) \diamond i(Mt) \diamond g(EH_{CC}) \diamond CC, o(Ct) \diamond d(EH_{CC}) \diamond CC, i(Fv) \diamond i(Ms) \diamond g(EH_{CC}) \diamond CC, o(Cs) \diamond d(EH_{CC}) \diamond CC \rightarrow CC \diamond i(Fa) \diamond i(Mt) \diamond g(EH_{CC}), CC \diamond o(Ct) \diamond d(EH_{CC}), CC \diamond i(Fv) \diamond i(Ms) \diamond g(EH_{CC}), CC \diamond o(Cs) \diamond d(EH_{CC})$

$TS_{EL} = i(Ct) \diamond i(Cs) \diamond i(Fa) \diamond i(Fv) \diamond g(EH_{EL}) \diamond EL, o(Cp) \diamond d(EH_{EL}) \diamond EL \rightarrow EL \diamond i(Ct) \diamond i(Cs) \diamond i(Fa) \diamond i(Fv) \diamond g(EH_{EL}), EL \diamond o(Cp) \diamond d(EH_{EL})$

The rules TS_{SC} denoted the saliency computation of the attention mechanism in thalamic-cortical circuits. Thalamic-cortical projection is an important infrastructure of brain function, and the thalamus plays an important role in the attention mechanism. Selective attention can reduce the influence of curse of dimensionality by saliency mechanism. In order to realize the saliency computation, this process focuses on the attention mechanism of the thalamic-cortical circuit and establishes the scheme of the saliency feature extraction. This rule mapping between the media MM and the spatial-temporal saliency features $\langle Sa, Sv \rangle$ was indicated as follows Equation (8).

$$\begin{aligned}
 SC:MM &\rightarrow \langle Sa, Sv \rangle \\
 s.t. \quad \langle Sa, Sv \rangle &= \prod_{(Mn, Es)} \sigma_{Mn}(MM)
 \end{aligned} \quad (8)$$

The rules TS_{DL} denotes the senses feature learning based on the hierarchical structure of the cortex. A cortical column is the basic unit of cognitive function. The cortex cognitive function is deep learning algorithm research basis and inspires how to realize the target classification and recognition. We can emulate the processing mechanism of the multi-layers architecture of the cortical column, and design the hierarchical semantic classifier. The probability distribution of the spatial-temporal senses features $\langle Fa, Fv \rangle$ was computed with media objects saliency features as follows Equation (9).

$$\begin{aligned}
 DL:\{ \langle Sa, Ma \rangle \langle Sv, Mv \rangle \} &\rightarrow \langle Fa, Fv \rangle \\
 s.t. \quad DL &= F_n \left(F_{n-1} \left(F_{n-2} (\dots, F_2(F_1(x)), \dots) \right) \right)
 \end{aligned} \quad (9)$$

The rules TS_{CC} denoted perceptual features computation based on probabilistic cognition. The computation process of the perceptual feature is building the mapping between Bayesian probability distribution of spatial-temporal senses features $\langle Fa, Fv \rangle$ and spatial-temporal perceptual features $\langle Ct, Cs \rangle$ as follows Equation (10).

$$\begin{aligned}
 CC:\{ \langle Fa, Mt \rangle; \langle Fv, Ms \rangle \} &\rightarrow \{ Ct; Cs \} \\
 s.t. \quad argmax_{\{Ct; Cs\}} P(\{ \langle Fa, Mt \rangle; \langle Fv, Ms \rangle \} | \{ Ct; Cs \}) \\
 &= \frac{P(\{ \langle Fa, Mt \rangle; \langle Fv, Ms \rangle \} | \{ Ct; Cs \}) P(\{ Ct; Cs \})}{P(\{ \langle Fa, Mt \rangle; \langle Fv, Ms \rangle \})}
 \end{aligned} \quad (10)$$

The rules TS_{EL} denoted target recognition based on multi-modal perception integration. It realizes the Ensemble Learning (EL) of multi-modal perception information and the final decision making of target semantic recognition. The core mission of target recognition is to establish the mapping between spatial-temporal senses-perceptual features $\langle Ct, Cs, Fa, Cp, Fv \rangle$ and target semantic labels as follows Equation (11).

$$\begin{aligned}
 EL:\{ Cs, Ct, Fv, Fa | DL, CC \} &\rightarrow Cp \\
 s.t. \quad Cp &= sign(\sum_m w_m f_m(\{ Cs, Ct, Fv, Fa | DL, CC \}))
 \end{aligned} \quad (11)$$

The rules TS_{RL} denoted the reward and punishment of emotion in the limbic system. It is the Reinforcement Learning (RL) basis that the emotions control of reward and punishment in the limbic system. The aim of simulating emotion control of rewards and punishment is to establish a stable and optimal target semantic. This rule solves errors minimization paradigm between the target semantic expectation Cp and the saliency feedback (Ei and Es) was defined as as follows Equation (12).

$$\begin{aligned}
 argMin_{\|RL\|} (\|Cp_L - Cp\|) \quad s.t. \quad RL: \{ Cp | CC \} \rightarrow \langle Ei, Es \rangle
 \end{aligned} \quad (12)$$

The rules TS_{IL} denoted the control of the memory system. The essence of semantic mapping is the memory and prediction for the spatial-temporal pattern. The material base for intelligent prediction includes the memory processing architecture of cortex-hippocampus circuits and its spatial-temporal pattern. This rule imitated mechanism of memory control, and storage and prediction of the historical information. The rules employed Incremental Learning (IL) method to control incremental knowledge. It includes the incremental of DNN's spatial-temporal features $\langle Mt, Ms \rangle$, the incremental of cognitive topic spatial-temporal features $\langle Mt, Ms \rangle$, and memory feedback Mp of incremental learning as follows Equation (13).

$$\begin{aligned}
 IL:\{ Ei, Cp | DL, CC \} &\rightarrow \langle Mp, Ma, Mv, Mt, Ms, Mn \rangle \\
 s.t. \quad M_t &= \frac{\alpha M_{t-1} Ei}{1 + sgn(\Delta t) exp(\gamma \Delta t sgn(\Delta t))} + \beta Cp, \\
 M_t &= M_{p_t}, M_{a_t}, M_{v_t}, M_{t_t}, M_{s_t}, M_{n_t}
 \end{aligned} \quad (13)$$

H. Algorithms Description for Semantic Learning and Recognition of Hierarchical CCNC Framework

The hierarchical CCNC framework semantic learning algorithm is dynamic process. It includes the following 8 steps as follows:

1:	It achieves spatial-temporal saliency features computation based on SNN according to the rules of TS_{SC} . $SS_{SC} // SS_{DL} // SS_{CC} // SS_{EL} // SS_{RL} // SS_{IL} \rightarrow SM_{SC} // SS_{DL} // SS_{CC} // SS_{EL} // SS_{RL} // SS_{IL}$
2:	It achieves target semantic learning of hierarchically integrated cognition based on deep learning and cognitive computing, including 3 dynamic processes as follows:
3:	It realizes spatial-temporal senses features computation of DNN according to the rules of TS_{DL} . $SM_{SC} // SS_{DL} // SS_{CC} // SS_{EL} // SS_{RL} // SS_{IL} \rightarrow SM_{SC} // SM_{DL} // SS_{CC} // SS_{EL} // SS_{RL} // SS_{IL}$
4:	It realizes spatial-temporal perception features computation of hierarchical topic model according to the rules of TS_{CC} . $SM_{SC} // SM_{DL} // SS_{CC} // SS_{EL} // SS_{RL} // SS_{IL} \rightarrow SM_{SC} // SM_{DL} // SM_{CC} // SS_{EL} // SS_{RL} // SS_{IL}$
5:	It realizes ensemble learning of objects semantic labels based on ensemble learning (such as AdaBoost et.al.) according to the rules of TS_{EL} . $SM_{SC} // SM_{DL} // SM_{CC} // SS_{EL} // SS_{RL} // SS_{IL} \rightarrow SM_{SC} // SM_{DL} // SM_{CC} // SM_{EL} // SS_{RL} // SS_{IL}$
6:	It achieves incremental computation and feedback of reinforcement learning based on object-oriented and multi-scale, including 2 dynamic processes as follows:
7:	It realizes multi-scale feedback computation of hierarchy reinforcement learning according to the rules of TS_{RL} . $SM_{SC} // SM_{DL} // SM_{CC} // SM_{EL} // SS_{RL} // SS_{IL} \rightarrow SM_{SC} // SM_{DL} // SM_{CC} // SM_{EL} // SM_{RL} // SS_{IL}$
8:	It realizes the temporal-spatial computation of object-oriented target based on online incremental learning according to the rules of TS_{IL} . $SM_{SC} // SM_{DL} // SM_{CC} // SM_{EL} // SM_{RL} // SS_{IL} \rightarrow SM_{SC} // SM_{DL} // SM_{CC} // SM_{EL} // SM_{RL} // SM_{IL}$

The hierarchical CCNC framework semantic recognition algorithm is also dynamic process. It includes the following 5 steps as follows:

1:	It achieves the saliency feature computation of sparse representation of SNN. $SS_{sc} // SS_{dl} // SS_{cc} // SS_{el} // SM_{rl} // SM_{il} \rightarrow SM_{sc} // SS_{dl} // SS_{cc} // SS_{el} // SM_{rl} // SM_{il}$
2:	It achieves target recognition of hierarchically integrated cognition based on deep learning and cognitive computing, including 3 processes as follows:
3:	It realizes spatial-temporal senses feature computation of DNN according to the rules of TS_{dl} . $SM_{sc} // SS_{dl} // SS_{cc} // SS_{el} // SM_{rl} // SM_{il} \rightarrow SM_{sc} // SM_{dl} // SS_{cc} // SS_{el} // SM_{rl} // SM_{il}$
4:	It realizes spatial-temporal perception features computation of hierarchical topic model according to the rules of TS_{cc} . $SM_{sc} // SM_{dl} // SS_{cc} // SS_{el} // SM_{rl} // SM_{il} \rightarrow SM_{sc} // SM_{dl} // SM_{cc} // SS_{el} // SM_{rl} // SM_{il}$
5:	It realizes ensemble computation of objects semantic labels based on ensemble learning (such as AdaBoost et.al.) according to the rules of TS_{el} . $SM_{sc} // SM_{dl} // SM_{cc} // SS_{el} // SM_{rl} // SM_{il} \rightarrow SM_{sc} // SM_{dl} // SM_{cc} // SM_{el} // SM_{rl} // SM_{il}$

REFERENCES

- [1] G.-W. Ng, *Brain-mind machinery: Brain-inspired computing and mind opening*, 2009.
- [2] Y. ZENG, et al., "Retrospect and Outlook of Brain-Inspired Intelligence Research," *Chinese Journal of Computers*, vol. 39, pp. 212-222, 2016. doi: <https://doi.org/10.11897/SP.J.1016.2016.00212>
- [3] D. S. Modha, et al., "Cognitive Computing," *Communications of the Acm*, vol. 54, pp. 62-71, Aug 2011. doi: <https://doi.org/10.1145/1978542.1978559>
- [4] A. Vestrucci, et al., "Can AI Help Us to Understand Belief? Sources, Advances, Limits, and Future Directions," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, p. 24, 2021. doi: <https://doi.org/10.9781/ijimai.2021.08.003>
- [5] Y. Chen, et al., "Neuromorphic computing's yesterday, today, and tomorrow – an evolutionary view," *Integration*, vol. 61, pp. 49-61, 2018. doi: <https://doi.org/10.1016/j.vlsi.2017.11.001>
- [6] E. R. Zhou, et al., "An improved memristor model for brain-inspired computing," *Chinese Physics B*, vol. 26, pp. 537-543, Nov 2017. doi: <https://doi.org/10.1088/1674-1056/26/11/118502>
- [7] C. Mead, "Neuromorphic electronic systems," presented at the Proceedings of the IEEE, 1990. doi: <https://doi.org/10.1109/5.58356>
- [8] K. Roy, et al., "Towards spike-based machine intelligence with neuromorphic computing," *Nature*, vol. 575, pp. 607-617, Nov 2019. doi: <https://doi.org/10.1038/s41586-019-1677-2>
- [9] Y. Wu, et al., "Spatio-Temporal Backpropagation for Training High-Performance Spiking Neural Networks," *Frontiers in Neuroscience*, vol. 12, pp. 1-12, 2018-May-23 2018. doi: <https://doi.org/10.3389/fnins.2018.00331>
- [10] Y. Sun, et al., "Quantum superposition inspired spiking neural network," *iScience*, vol. 24, 2021. doi: <https://doi.org/10.1016/j.isci.2021.102880>
- [11] B. V. Benjamin, et al., "Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations," *Proceedings of the IEEE*, vol. 102, pp. 699-716, 2014. doi: <https://doi.org/10.1109/jproc.2014.2313565>
- [12] S. Schmitt, et al., "Neuromorphic hardware in the loop: Training a deep spiking network on the BrainScaleS wafer-scale system," in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 2227-2234.
- [13] S. B. Furber, et al., "The SpiNNaker Project," *Proceedings of the IEEE*, vol. 102, pp. 652-665, 2014. doi: <https://doi.org/10.1109/jproc.2014.2304638>
- [14] D. Ma, et al., "Darwin: A neuromorphic hardware co-processor based on spiking neural networks," *Journal Of Systems Architecture*, vol. 77, pp. 43-51, Jun 2017. doi: <https://doi.org/10.1016/j.sysarc.2017.01.003>
- [15] J. Pei, et al., "Towards artificial general intelligence with hybrid Tianjic chip architecture," *Nature*, vol. 572, pp. 106-111, 2019/08/01 2019. doi: <https://doi.org/10.1038/s41586-019-1424-8>
- [16] M. Paul A, "Artificial brains. A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 6197, pp. 668-673, 2014. doi: <https://doi.org/10.1126/science.1254642>
- [17] N. P. Jouppi, et al., "Motivation for and Evaluation of the First Tensor Processing Unit," *IEEE Micro*, vol. 38, pp. 10-19, May-Jun 2018. doi: <https://doi.org/10.1109/MM.2018.032271057>
- [18] T. Luo, et al., "DaDianNao: A Neural Network Supercomputer," *Ieee Transactions on Computers*, vol. 66, pp. 73-88, Jan 2017. doi: <https://doi.org/10.1109/tc.2016.2574353>
- [19] B. J. Kagan, et al., "In vitro neurons learn and exhibit sentence when embodied in a simulated game-world," *Neuron*, vol. 110, pp. 1-18, December 7, 2022 2022. doi: <https://doi.org/10.1016/j.neuron.2022.09.001>
- [20] R. Wang, et al., "Neuromorphic Hardware Architecture Using the Neural Engineering Framework for Pattern Recognition," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 11, pp. 574-584, 2017. doi: <https://doi.org/10.1109/tbcas.2017.2666883>
- [21] M. V. DeBole, et al., "TrueNorth: Accelerating From Zero to 64 Million Neurons in 10 Years," *Computer*, vol. 52, pp. 20-29, 2019. doi: <https://doi.org/10.1109/MC.2019.2903009>
- [22] C. Eliasmith, et al., "A large-scale model of the functioning brain," *Science*, vol. 338, pp. 1202-1205, 2012. doi: <https://doi.org/10.1126/science.1225266>
- [23] A. Mehonic and A. J. Kenyon, "Brain-inspired computing needs a master plan," *Nature*, vol. 604, pp. 255-260, 2022/04/01 2022. doi: <https://doi.org/10.1038/s41586-021-04362-w>
- [24] G. Piccinini, "The First Computational Theory of Mind and Brain: A Close Look at McCulloch and Pitts's "Logical Calculus of Ideas Immanent in Nervous Activity"," *Synthese*, vol. 141, pp. 175-215, 2004/08/01 2004. doi: <https://doi.org/10.1023/B:SYNT.0000043018.52445.3e>
- [25] V. Mnih, et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529-533, 02/26/print 2015. doi: <https://doi.org/10.1038/nature14236>
- [26] D. Silver, et al., "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, pp. 484-489, 01/28/print 2016. doi: <https://doi.org/10.1038/nature16961>
- [27] S. Sabour, et al., "Dynamic Routing Between Capsules," presented at the neural information processing systems, 2017
- [28] A. Creswell, et al., "Generative Adversarial Networks: An Overview," *IEEE Signal Processing Magazine*, vol. 35, pp. 53-65, 2018. doi: <https://doi.org/10.1109/MSP.2017.2765202>
- [29] M. G. Huddar, et al., "Attention-based Multi-modal Sentiment Analysis and Emotion Detection in Conversation using RNN," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, pp. 112-121, 2021. doi: <https://doi.org/10.9781/ijimai.2020.07.004>
- [30] J. Bobadilla, et al., "DeepFair: Deep Learning for Improving Fairness in Recommender Systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, pp. 86-94, Jun 2021. doi: <https://doi.org/10.9781/ijimai.2020.11.001>
- [31] A. Vaswani, et al., "Attention Is All You Need," in *31st Annual Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, 2017, pp. 1-15.
- [32] T. B. Brown, et al., "Language Models are Few-Shot Learners," presented at the Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS 2020), Virtual-only Conference, 2020. doi: <https://doi.org/arXiv:2005.14165> [cs.CL]
- [33] Z. Wu, et al., "A Comprehensive Survey on Graph Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1-21, 2020. doi: <https://doi.org/10.1109/TNNLS.2020.2978386>
- [34] D. George and J. Hawkins, "Towards a Mathematical Theory of Cortical Micro-circuits," *Plos Computational Biology*, vol. 5, Oct 2009. doi: <https://doi.org/10.1371/journal.pcbi.1000532>
- [35] N. Tzourio-Mazoyer, et al., "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *Neuroimage*, vol. 15, pp. 273-289, 2002. doi: <https://doi.org/10.1006/nimg.2001.0978>
- [36] L. Fan, et al., "The Human Brainnetome Atlas: A New Brain Atlas Based on Connectonal Architecture," *Cerebral Cortex*, vol. 26, pp. 3508-3526, 2016. doi: <https://doi.org/10.1093/cercor/bhw157>
- [37] Y. Liu, et al., "Cognitive Neural Mechanisms and Saliency Computational Model of Auditory Selective Attention," *Computer Science*, vol. 40, pp. 283-287, 2013. doi: <https://doi.org/10.3969/j.issn.1002-137X.2013.06.063>

[38] Y. Liu, *et al.*, "Cognitive Neural Mechanisms and Saliency Computational Model of Visual Selective Attention," *Journal of Chinese Computer Systems*, vol. 35, pp. 584-589, 2014.[doi.https://doi.org/10.3969/j.issn.1000-1220.2014.03.029](https://doi.org/10.3969/j.issn.1000-1220.2014.03.029)

[39] X. Fu, *et al.*, "A computational cognition model of perception, memory, and judgment," *Science China-Information Sciences*, vol. 57, Mar 2014.<https://doi.org/10.1007/s11432-013-4911-9>

[40] P. T. Fox, *et al.*, "Meta-analysis in human neuroimaging: computational modeling of large-scale databases," *Neuroscience*, vol. 37, pp. 409-434, 2014.<https://doi.org/10.1146/annurev-neuro-062012-170320>

[41] E. C. Cieslik, *et al.*, "Is there "one" DLPFC in cognitive action control? Evidence for heterogeneity from co-activation-based parcellation," *Cerebral Cortex*, vol. 23, pp. 2677-2689, 2013.[doi.https://doi.org/10.1093/cercor/bhs256](https://doi.org/10.1093/cercor/bhs256)

[42] S. B. Eickhoff, *et al.*, "Co-activation patterns distinguish cortical modules, their connectivity and functional differentiation," *Neuroimage*, vol. 57, pp. 938-949, 2011.<https://doi.org/10.1016/j.neuroimage.2011.05.021>

[43] Y. Liu, *et al.*, "CSRNCVA: A model of cross-media semantic retrieval based on neural computing of visual and auditory sensations," *Neural Network World*, vol. 28, pp. 305-323, 2018.<https://doi.org/10.14311/NNW.2018.28.018>

[44] Y. Liu, *et al.*, "CSMCCVA: Framework of cross-modal semantic mapping based on cognitive computing of visual and auditory sensations," *High Technology Letters*, vol. 22, pp. 90-98, 2016.<https://doi.org/10.3772/j.issn.1006-6748.2016.01.013>

[45] J. S. Gero, "Design Prototypes: A Knowledge Representation Schema for Design," *AI Magazine*, vol. 11, pp. 26-36, 1990

[46] Y. Zhong, "Mechanism-based artificial intelligence theory: a universal theory of artificial intelligence," *CAAI Transactions on Intelligent Systems*, vol. 13, pp. 2-18, 2018.<https://doi.org/10.11992/tis.201711032>

[47] Y. Liu, *et al.*, "Scene classification of high-resolution remote sensing image based on multimedia neural cognitive computing," *Systems Engineering and Electronics*, vol. 37, pp. 2623-2633, 2015.<https://doi.org/10.3969/j.issn.1001-506X.2015.11.31>

[48] Y. Liu, *et al.*, "Spike-Based Approximate Backpropagation Algorithm of Brain-Inspired Deep SNN for Sonar Target Classification," *Computational Intelligence and Neuroscience*, vol. 2022, p. 1633946, 2022/10/20 2022.<https://doi.org/10.1155/2022/1633946>

[49] Y. Liu, *et al.*, "SAR ship detection using sea-land segmentation-based convolutional neural network," presented at the 2017 International Workshop on Remote Sensing with Intelligent Processing, RSIP 2017, May 19, 2017 - May 21, 2017, Shanghai, China, 2017.<https://doi.org/10.1109/RSIP.2017.7958806>

[50] Y. Liu, *et al.*, "Target detection in remote sensing image based on saliency computation of spiking neural network," presented at the 38th Annual IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2018, July 22, 2018 - July 27, 2018, Valencia, Spain, 2018.<https://doi.org/10.1109/IGARSS.2018.8517588>

[51] Y. Liu and F.-b. Zheng, "Object-oriented and multi-scale target classification and recognition based on hierarchical ensemble learning," *Computers & Electrical Engineering*, vol. 62, pp. 538-554, 2017.<https://doi.org/10.1016/j.compeleceng.2016.12.026>

[52] Y. Liu, *et al.*, "Hyperspectral Image Classification of Brain-Inspired Spiking Neural Network Based on Approximate Derivative Algorithm," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-16, 2022.<https://doi.org/10.1109/tgrs.2022.3207098>

[53] Y. Liu, *et al.*, "Research of Neural Cognitive Computing Model for Visual and Auditory Cross-media Retrieval," *Computer Science*, vol. 42, pp. 19-25, 30, 2015 2015.<https://doi.org/10.11896/j.issn.1002-137X.2015.3.004>

[54] C. Pogliano, "Lucky Triune Brain Chronicles of Paul D-MacLean's Neuro-Catchword," *Nuncius-Journal of the History of Science*, vol. 32, pp. 330-375, 2017.<https://doi.org/10.1163/18253911-03202004>

[55] S. CU, "The triune brain in antiquity: Plato, Aristotle, Erasistratus," *Journal of the history of the neurosciences*, vol. 1, pp. 1-14, 2010.<https://doi.org/10.1080/09647040802601605>

[56] R. Joseph, *Neuropsychiatry, Neuropsychology, Clinical Neuroscience*, 2000.

[57] T. L. Griffiths, *et al.*, *Bayesian models of cognition*: Cambridge University Press, 2008.

[58] A. A. Valladares, *et al.*, "Event related potentials changes associated with the processing of auditory valid and invalid targets as a function of

previous trial validity in a Posner's paradigm," *Neuroscience Research*, vol. 115, pp. 37-43, Feb 2017.<https://doi.org/10.1016/j.neures.2016.09.006>

[59] G. Berry and G. Boudol, "The chemical abstract machine," *Theoretical Computer Science*, vol. 96, pp. 217-248, 1992/04/06/ 1992.[https://doi.org/10.1016/0304-3975\(92\)90185-I](https://doi.org/10.1016/0304-3975(92)90185-I)

[60] X. Li, *et al.*, "A New Method to Construct the Software Vulnerability Model," presented at the 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), Beijing, 2017.<https://doi.org/10.1109/CIAPP.2017.8167212>



Yang Liu

Yang Liu (Member of IEEE, CCF and CAAI) received the B.S. degree from Changchun University of Science and Technology in 1996, the M.S. degree and Ph.D. degree from Henan University in 2009 and 2016, respectively. He is currently a Professor and Ph.D. supervisor of the College of Computer Science and Information Engineering, Henan University. He is the Principle Investigator of Brain-inspired Intelligence Science and Technology Innovative Team. His research interests include scientific theory of Brain&Mind-inspired Computing (BMC), common technology of multimedia information processing, and engineering application of spatiotemporal big data. Specifically, he focuses on exploring the models of mind-like computing such as multimedia neural cognitive computing, and cross-modal cognitive neural computing. Currently, he mainly researches the key algorithms of brain-inspired computing such as cross-modal target recognition, cross-media semantic retrieval, multi-source and multimedia information processing. He has also developed the complex systems of high performance computing such as spatiotemporal information processing platform, and remote sensing intelligent information extraction development kit.



Jianshe Wei

Jianshe Wei is Distinguished Professor, Yellow River scholar and doctoral supervisor of School of Life Sciences, as Principle Investigator of Institute for Brain Sciences Research in Henan University. He received Medicine Degree from Xinxiang Medical College in 1993 and Ph.D. degree from Fudan University in 2002. He worked as a postdoctoral researcher at the University of Calgary, the Tokyo Metropolitan Institute for Neuroscience and New York State Institute for Basic Research in Developmental Disabilities from 2003 to 2010. The main research direction focuses the basic problems of neurodegenerative diseases in Parkinson's disease by molecular pathology and biophysical methods. At present, it is mainly engaged in the basic research of neural circuit and synaptic plasticity in affective disorders.

Deep Transfer Learning-Based Automated Identification of Bird Song

Nabanita Das^{1*}, Neelamadhab Padhy¹, Nilanjan Dey², Sudipta Bhattacharya³, João Manuel R.S. Tavares⁴

¹ Department of Computer Science and Engineering, GIET University, Gunupur (India)

² Department of Computer Science & Engineering, Techno International New Town, Kolkata (India)

³ Department of Computer Science & Engineering, Bengal Institute of Technology, Kolkata (India)

⁴ Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Porto (Portugal)

Received 13 September 2022 | Accepted 14 December 2022 | Published 12 January 2023



ABSTRACT

Bird species identification is becoming increasingly crucial for avian biodiversity conservation and assisting ornithologists in quantifying the presence of birds in a given area. Convolutional Neural Networks (CNNs) are advanced deep learning algorithms that have proven to perform well in speech classification. However, developing an accurate deep learning classifier requires a large amount of data. Such a large amount of data on endemic or endangered creatures is frequently difficult to gathered. Also, in some other fields, such as bioinformatics and robotics, the high cost of data collection and expensive annotation limit their progress, so large, well-annotated data creating a set is also difficult. A transfer learning method can alleviate overfitting concerns in a deep learning model. This feature serves as the inspiration for transfer learning, which was created to deal with situations where the data are distributed across a variety of functional domains. In this study, the ability of deep transfer models such as VGG16, VGG19 and InceptionV3 to effectively extract and discriminate speech signals from different species of birds with high prediction accuracy is explored. The obtained accuracies using VGG16, VGG19 and InceptionV3 were equal to 78, 61.9 and 85%, respectively, which are very promising.

KEYWORDS

Bird Species Recognition, Convolution Neural Network, Data Augmentation, InceptionV3, Transfer Learning, VGG16, VGG19.

DOI: 10.9781/ijimai.2023.01.003

I. INTRODUCTION

BIRDS not only enhance nature's charm and beauty but also help maintain the balance of the new environment of the world. Because they are essential parts of natural systems, birds have ecological importance. Birds manage insects and rodents, pollinate crops, spread seeds, and serve humans directly. Bird vocalizations are very noticeable, which makes them a helpful tool for population monitoring and biodiversity assessment. Bird vocalization includes both calls and songs. Birds are essential to our ecology. For instance, birds keep our globe beautiful by controlling pests, pollinating crops, and preserving the ecology of an island. There are around 10,000 species on earth, according to [1]. Birds make sounds for many reasons, including locating territories, which is important for male birds, inviting a mate to mate, reacting to their environment, and determining whether or not they are in danger [2]. People often find it difficult to distinguish between a bird's song and a call, especially if they are unfamiliar with birds. An audio recording of a bird's voice is an essential tool for identifying the species of a bird for a biologist who is interested in the study, management, and conservation of birdlife [3]. There are many bird calls, and it is hard for people to figure out

which ones have a place with animal categories. The manual recording and recognition of avian sounds is inconvenient and can sabotage bird conservation efforts. As a result, accurate, scalable, and automated bird species recognition is essential for wildlife monitoring and can help conserve avian biodiversity [4]-[7]. The identification of bird species is a classic pattern recognition problem, and most research includes sections on signal pre-processing, feature extraction, and classification [8], [9]. Deep learning has received increased attention from researchers recently since it has been successfully used in a number of practical applications. To stop the rapid loss of avian variety in this area, deep-learning algorithms for bird detection are appropriate [10]. In this context, several automated bird detection models were developed. Additionally, a test has been performed on a system that can recognize new bird songs and learn from previously recorded annotated bird sounds. This system may provide accurate information on the presence or absence of a target species as well as the overall biodiversity status of a region. Deep learning algorithms are superior to conventional machine learning techniques because they can extract high-level characteristics from enormous amounts of data [11]. On the other hand, traditional machine learning approaches need users to construct features, which demands significant manual work.

On the other hand, deep learning approaches automatically extract data features using a hierarchical feature extraction method and an unsupervised or semi-supervised feature learning methodology [12]-[14]. Deep learning can be defined as a representation learning

* Corresponding author.

E-mail address: nabanita.das@giuet.edu

algorithm in machine learning that is based on large data. Although deep learning models can achieve good predictive performance, such models require a huge number of unique data points to achieve this performance and this turns out to be challenging for endangered or endemic birds, as inadequate data overwhelms deep learning models. One of the fundamental problems of deep learning is data dependency. Deep learning is more dependent on training data than traditional machine learning methods since it needs a lot of data to find latent patterns in the data. Inadequate training data is unavoidable in some deep-learning applications. For instance, the high cost of data collection and expensive annotation, which impedes development, make it difficult to produce a sizable, thoroughly annotated dataset for each sample in a bioinformatics dataset [15], [16],[17]. The issue of overfitting in a deep model can be solved using a transfer learning technique [18]. Because transfer training makes the condition that the training data be independent and distributed equally with the test data simpler, it can address the issue of a lack of training data. Transfer learning drastically reduces the amount of training data and time needed for the target domain, because it does not require training and testing data or starting from scratch to train the target domain model.

In this experimental study, 7 different bird species were correctly identified using 16387 test samples from the xeno-canto database. Due to the limited sample size, several data augmentation techniques have been studied, and the underlying hypotheses were thoroughly evaluated. It was interesting to note that such type of augmentation techniques results in overfitting of the models. In light of these considerations, the idea of transfer learning was chosen for this investigation. By using transfer learning, a total of 36 species were classified rather than 7. Because of the limited availability of high-quality data, pre-trained models have been used in the identification of 37 different categories of birds. To develop this investigation on the local bird recognition in Sundarban, West Bengal, India, two deep learning models were used. Hence, InceptionV3 and MobileNet were initially tested without the use of transfer learning technology, and then they were tested once again with it. Finally, the findings were compared using MobileNet and transfer learning, employing performance evaluation metrics such as accuracy and F1 score. In the experiment, the result showed that in the VGG16 model the training accuracy was 75%, while the test one accuracy was 78%. Respectively in the VGG19 model, the training accuracy was 64%, while the test accuracy was 61.9%. On the InceptionV3 model, which was employed in additional tests, an accuracy of 95% was reached during training, while an accuracy of 85% was achieved which obtained the best result. In the InceptionV3 model, ImageNet was used as a weight, and average pooling was employed. The rest of this article is organized in the following way. Section 2 describes the related research. The methodologies used are detailed in Section 3. Section 4 shows the results and their analysis, which are followed by a discussion and conclusion in section 5, and concludes what future work can be done.

II. LITERATURE SURVEY

Different researchers have proposed different features for the audio sounds of birds, and artificial intelligence techniques have been used to voice classification. CNN models that use Mel spectrogram or mel frequency cepstral coefficient (MFCC) derived from audio data have been observed to dominate the most promising solutions [19]. However, recent trends show that the best results were achieved by the works that used Convolutional Neural Networks with transfer learning [20]. The best results were for the most part from using Resnet, Inception, and VGG models. Additionally, Fritzler et al. [21] propose the Inception-v3 pre-trained convolutional neural network-based bird recognition system. The technology was enhanced with 36,492 audio

recordings of 1,500 different bird species for the BirdCLEF 2017 task. The audio recordings were afterward transformed into spectrograms and used for data augmentation. According to this study, optimizing a pre-trained convolutional neural network trumps starting from scratch in terms of performance. For acoustic bird detection, Ntalampiras [22] introduced a transfer learning framework employing the probability density distribution of ten musical genres to determine the degree of affinities between different bird species and various musical genres. Deep learning models based on CNNs are efficient categorization models. However, getting numerous training samples in specialist disciplines like bird acoustics is expensive and difficult as they require a large amount of data for training. To address this issue, transfer learning is one method that can classify data with a limited number of training examples. DB Efremov et al. [23] assessed the effectiveness of birdcall classification utilizing transfer learning from a bigger base dataset to a smaller target dataset using a ResNet-50 CNN in this regard. A bird recognition model built on Inception-v3 was presented by J. Bai et al. [24] can identify and categorize 659 different bird species from supplied audio recordings. Inception-v3 is used to recognize bird sounds by using log-Mel spectrograms as features.

To enhance the model's performance, several data augmentation strategies were employed. In order to categorize the cries of 24 species of birds and amphibians discovered in environmental field recordings, Zhong M. et al. [25] created a deep convolutional neural network. Their primary objective was to prepare enough training data, which is a significant difficulty for many deep-learning applications. To tackle this problem, they created a pre-trained deep convolutional neural network by fusing the idea of transfer learning with a supervised pseudo-labeling technique and an eigen loss function. In order to categorize grouper species based on the courtship-related noises they make during spawning aggregations, Ibrahim suggests a transfer learning technique, A. K. [26]. On the other hand, Rajan R. et al. [27] suggested a method for learning bird vocalizations utilizing sliding window analysis on the Mel spectrogram and a pre-trained Deep Convolutional Neural Network (DCNN), a VGG16 model. Using a deep learning model, Henri, E. J. et al. [28] created a method for classifying Mauritius bird sounds from audio recordings. Many categorized recordings from the birdsong-sharing website Xeno-canto were utilized as input for this model. Following that, they improved three previously trained CNN models: InceptionV3, MobileNetV2, and ResNet50, as well as a brand-new model. With 84% of accuracy, transfer learning was successfully applied to develop the study's model. However, to create an effective deep-learning classifier, a substantial amount of data is needed. It is typically difficult to gather such vast amounts of data about endemic or endangered organisms. By separating two acoustic features, mainly, the Mel spectrogram and the Mel frequency cepstral coefficient, from each data point, Gunawan, K.W. et al. [29] established a transfer learning model that restricts overfitting in deep models and a method to maximize the dataset used. In order to incorporate and learn from both audio data, the researchers employed a two-input scalable convolutional neural network constructed from EfficientNet. On the test set, they had 99.9% of accuracy. A classification system for the sounds of 17 species of Indian owls was developed by Nayak S. et al. [30]. For the transfer learning model created in this study, four model architectures were used: InceptionV3, Resnet152, InceptionResnetV2, and VGG16, with all models sharing the same model parameters. The InceptionV3 network, which had an accuracy of 85.3%, produced the most precise results. ResNet50, DenseNet201, InceptionV3, Xception, and Efficient Net were just a few of the deep transfer learning models employed by Kumar Y. et al. [31] to create an intelligent system for predicting various bird species from a massive collection of audio data sets. DenseNet201 has the highest classification accuracy in the group, which was of 97.43%. A methodology for automatically classifying and

TABLE I. STATE OF ART STUDIES ON AUTOMATED BIOACOUSTICS BIRD SPECIES IDENTIFICATION

Author(s)	Dataset(s)	Technique(s)	Limitation(s)	Results
Sprengel, E. et al. (2016)	LifeCLEF plant challenge 2016 Dataset	CNN	Longer files create chunks	Accuracy: 84%
M Lasseck (2018)	LifeCLEF 2018	DCNNs pre-trained on ImageNet	Results can be further enhanced by combining models with various features	Accuracy: 93%
Ntalampiras, S. (2018)	GTZAN corpus and http://www.Xeno-canto.org/	Transformation based on Reservoir Networks	It is necessary to assess whether the Transfer Learning-based approach can handle feature spaces with a wide range of sizes	There were obtained 92.5 and 81.3% classification accuracy on average
Efremova et al. (2019)	From http://www.Xeno-canto.org/ : Base "SoundNet" Dataset, Target Dataset, Negative Dataset	ResNet-50 CNN	Results can be further improved	In 5-fold cross-validation, the target dataset's average validation accuracy was of 79%
Bai, J., et al. (2019)	BirdCLEF2019	Inception-v3	Ensemble of networks could significantly improve the results	The classifications mean average precision was of 0.055 (c-mAP)
Rahman, M. M., et al. (2020)	Seven local birds' images	MobileNet and Inception-v3	Need to evaluate whether this model is suitable for a large number of various species	Accuracy: 91%
R Rajan., et al. (2021)	Xeno-canto bird sound database	VGG16 through a sliding window analysis on Mel spectrogram	The classification of multiple-label birds is a difficult undertaking because of vocalization that overlaps	Average F1-score: 0.65
Henri, E. J., et al. (2021)	Xeno-canto bird sound database	InceptionV3, MobileNetV2 and ResNet50	Misclassifications were detected in some classes	Accuracy: 84%
Gunawan, K. W., et al. (2021)	Xeno-canto database	Scalable with two inputs, EfficientNet's Convolutional Neural Network (CNN)	It is difficult to gather the vast amount of high-quality data on endemic or threatened animals that are required to create a powerful model	Accuracy: 99.27%
Nayak, S., et al. (2022)	Xeno-canto database	The ImageNet dataset was used to train the pre-trained InceptionV3 network	Need to detect the calls in poor quality audio	Accuracy: 85.3%
Kumar, Y., et al. (2022)	https://www.kaggle.com/c/birdsongrecognition/data	InceptionV3, Xception, ResNet50, DenseNet201, and Efficient Net	To increase the recognition rate, various noise reduction filtering must be applied during the pre-processing stage	DenseNet201 and ResNet50 classification models achieved an accuracy of 97.43% on the validation set.
Sharma, N., et al. (2022)	With 264 bird species, Cornell Bird Call Identification - 200 dataset offers roughly 150 recordings for each one	ResNet50V2 and EfficientNetB0	Need to detect the calls in poor quality audio and need to remove ambient noise	EfficientNetB0 accuracy: 92.4%

processing images and sounds to identify bird species from bird videos was presented by Sharma N. et al. [32]. On image and sound datasets containing recordings of 137 different bird species, classification models for images and sounds were developed using pre-trained neural networks ResNet 50V2 and EfficientNet B0. The final model's overall accuracy was equal to 90%, while the test accuracy for the two models was 97.1 and 92.4%, respectively.

Deep learning techniques would be a practical solution, according to the aforementioned discussion of previously developed methodologies [33], [34]. The creation of a useful classification model that optimizes performance for numerous species using transfer learning and convolutional neural networks is the major contribution of the current study. Ornithologists and other researchers are aware of the potential benefits that may come from combining developments in bioacoustics with transfer learning models, which could provide a new study dimension. Additionally, there have been a few works completed, some of which we have discussed here; nonetheless, all of

those efforts have certain restrictions. Table I lists state of art studies on automated bioacoustics bird species identification by using transfer learning models. It has been noted that exceptionally lengthy files may sometimes break apart into pieces. It is essential to determine whether or not the Transfer Learning-based strategy can manage feature spaces that span a broad range of sizes. Following the deployment of transfer learning models, it was shown that the categorization of multiple-label birds might be a challenging task at times due to overlaps in their vocalizations. Additionally, misclassifications were found in certain classes. Utilizing transfer learning models has not resulted in a significant amount of additional work being done for the purpose of recognizing calls from low-quality audio. During the pre-processing step, a number of noise-reduction filtering techniques need to be employed in order to get a higher recognition rate. Our main focus of this work is to provide a technique that can identify a large number of species from their audio and also the system must be cost-effective and scalable.

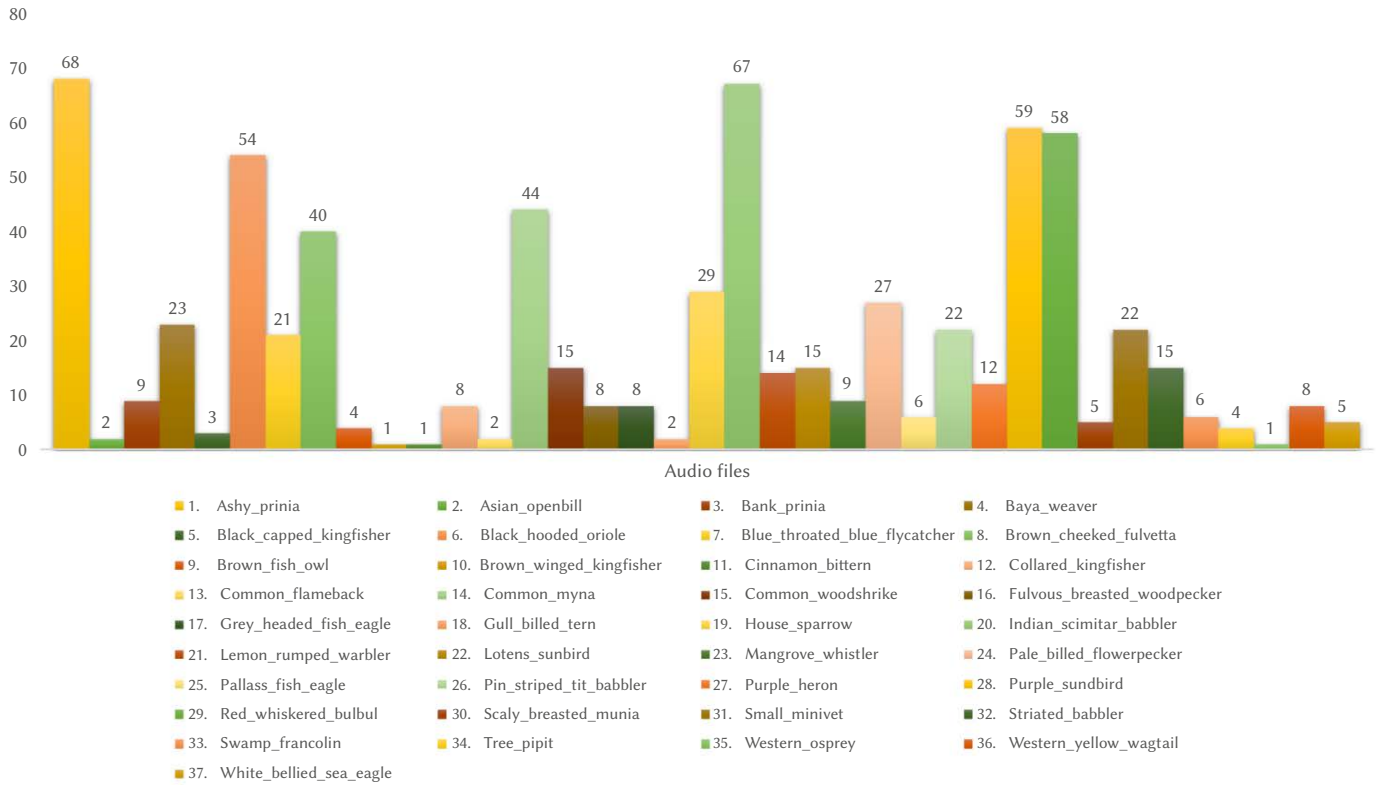


Fig. 1. List of the studied species data.

III. METHODOLOGY

This section outlines the procedure followed in this study. The employed methodology incorporates transfer learning, deep learning, and audio-processing ideas. First, input comes from an audio recording of the bird under analysis. After that, features are extracted from the audio input using signal pre-processing techniques. The processed components are then fed into a powerful classification model that makes use of Convolutional Neural Networks [35], [36] and the idea of Transfer Learning [37], [38] to produce the best results for a wide range of species. The used three models are built on pre-trained networks called VGG16, VGG19, and InceptionV3, which were trained using data from 37 different bird species. In Fig. 2, the implementation process for this study is shown. First, from Xeno Canto, particular regional data is chosen, then data cleaning is performed, and after that data augmentation technique is used. Then, all the data is inputted into pre-trained models and classified.

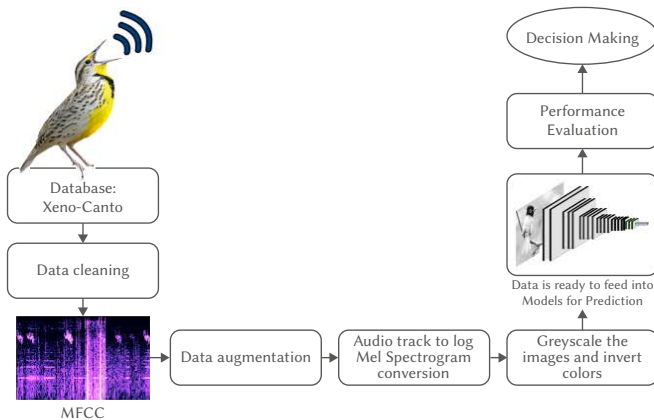


Fig. 2. Proposed bird sound identification solution.

A. Data Collection

The widely used Xeno-canto bird sound database served as the foundation for this study's dataset. Volunteers from all across the world can record bird calls and sounds for the Xeno-canto Foundation, an online database of bird noises that includes more than a million bird sounds from more than 10,000 distinct species. Birdsong captured at Sundarban, West Bengal, India, served as the particular dataset for this study. Information on 37 different species was gathered. Fig. 1 shows the dataset utilized in this study. Each audio file was modified to contain a single vocalization lasting 1.5 seconds (sampling rate: 16000 Hz). In total, 11325 files were included. The models were trained with augmented data, which were validated using the original 453 files.

B. Data Cleaning

Data cleaning is the process of removing inaccurate, corrupted, malformed, duplicate, or incomplete data from a dataset. There is a substantial risk of data duplication or mislabeling when merging multiple data sources. Background noise in the downloaded audio files was minor, which was confirmed manually. Hence background noise treatment was unnecessary. Parts of the audio files that had no or minimal sound were eliminated as follows: firstly, it was determined what the median sound power was, and the audio segments whose energy level or functional ability was below 50% of the median were removed, and lastly, the remaining audio files were reassembled.

C. Feature Extraction Technique

1. Mel Spectrogram

The audio sample was converted to Mel spectrogram in a different way and at different frequencies. Humans always perceive frequency logarithmically. A time-frequency representation, a perceptually appropriate amplitude representation, and ultimately a perceptually

relevant frequency representation make up ideal sound qualities. For the pitch, Mel is crucial. Convert the frequencies to the Mel scale, extract the short-time Fourier transform, and then convert the amplitude to Db.

The Mel scale conversion procedures for frequencies are:

- Determine the number of Mel Scales;
- Create banks of Mel filters;
- Use Mel filter banks for the spectrogram.

D. Data Set Pre-Processing

Data pre-processing is the first and most crucial stage in developing a classification model. The audio classification task is an image classification challenge in this study. Here, MFCCs are employed in sound identification tasks and can accurately map auditory information in a visual domain (Fig. 3.). In order to be used, CNN models for classification audio recordings must be represented in the optical environment. Different processing is frequently required to make the dataset acceptable for usage with a CNN model [39]. The data pre-processing steps include data sizing, labeling, and expansion. The database consists of audio of the 37 birds' songs of Sundarban; among them, 19 birds' themes are included, which are very few (below 10). This significant imbalance may influence the model's performance and can lead to issues such as overfitting and difficulty learning the model.

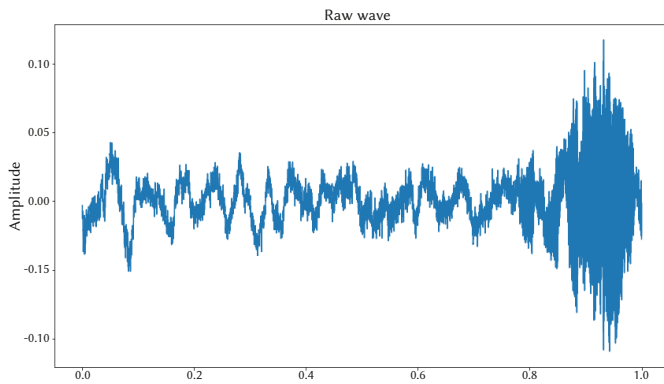


Fig. 3. Time domain representation of original audio of the ashy_prina dataset.

1. Data Augmentation

A CNN model could not be used since the data collected for certain species was insufficient. Therefore, for those specific species, data augmentation was used. On the other hand, in order to prevent overfitting, data augmentation is needed. The term "data augmentation" refers to an increase in available data. Time shifting, adding noise, time stretching, and pitch augmentation is examples of audio data augmentation techniques. Time stretching, pitch scaling, and the addition of white noise were the three data augmentation methods used in this study. The aforementioned data-cleaning procedure has been applied to all used data.

a) Time Stretching

A method is known as "time stretching" allows one to increase the length or speed of an audio stream without changing its pitch or other parameters. For example, one can extend a sound to 200 milliseconds by decoding twice as many samples from each frame if uttered for 100 milliseconds (10 frames) [40]. Librosa, a python utility for music modification, applies the time stretching simple. The rate settings can change the audio's pace and duration. Fig. 4 represents the time stretching of 0.8 times of original audio.

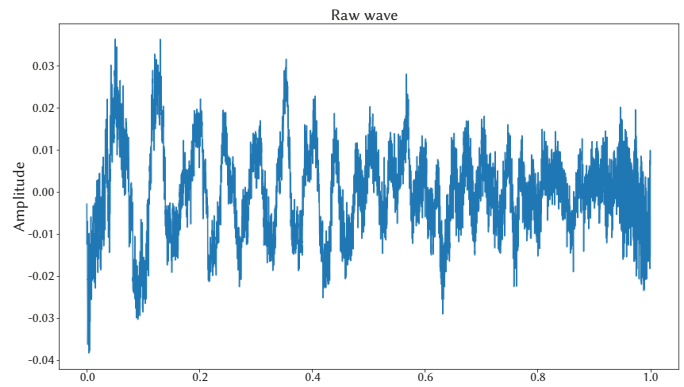


Fig. 4. Time Stretching of an original bird call audio.

b) Pitch Scaling

This technique serves as a wrapper for the librosa function. The pitch veers all over the place. When applying different rate values without altering the duration of the signal, pitch scaling is the reverse of time stretching [41], as can be seen in Fig. 5.

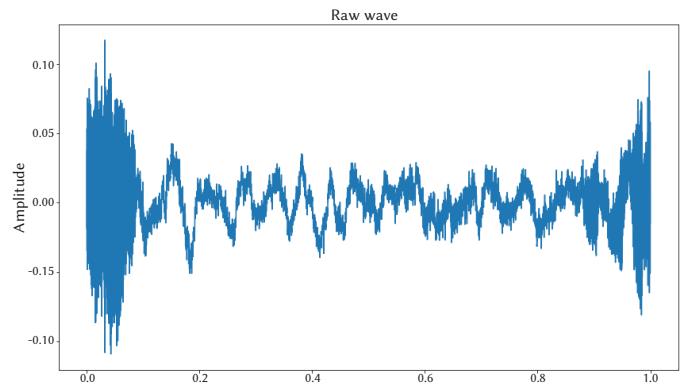


Fig. 5. Pitch Scaling of an original bird call audio.

c) Noise Addition

Noise addition can generate syntactic audio data for the data augmentation process. Numpy makes it simple to deal with noise addition by adding a random value to the date. In Fig. 6, this technique can be seen.

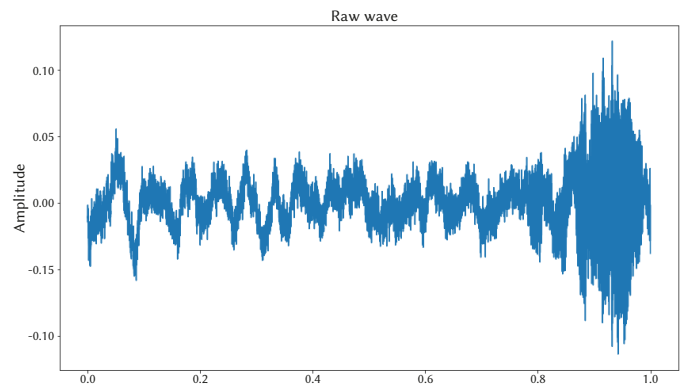


Fig. 6. Noise addition of an original bird call audio.

2. Dataset Splitting: Training & Testing

The total number of files that were selected from Sundarbans's set consisted of 2265; after doing data augmentation, the total number of files was 11325. Then, the used dataset was split into 80% and 20%, for training and testing, respectively.

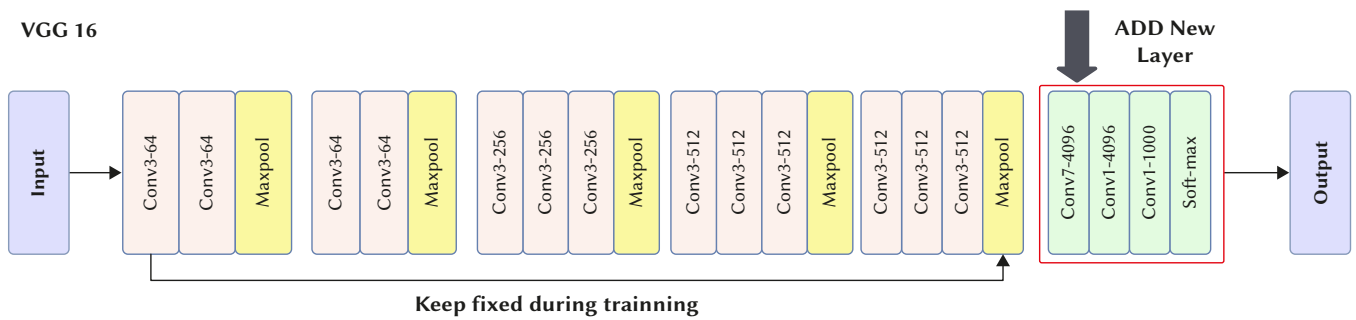


Fig. 7. VGG16 model architecture.

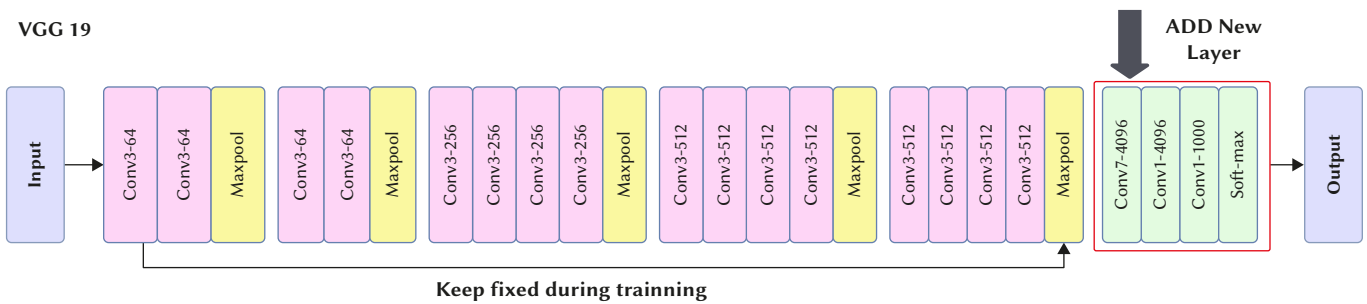


Fig. 8. VGG19 model architecture.

E. Model Description

The use of deep learning in audio recognition is well-recognized. Neural networks have been applied to numerous facets of audio recognition since the development of deep understanding [42], [43]. The effectiveness of neural learning for sound recognition is influenced by the adaptability and predictive power of the increasingly accessible deep neural networks. The deep learning models utilized in the study are described in the following.

1. VGG16

VGGNet-16 has a relatively homogeneous architecture with 16 convolutional layers. It only has 3x3 convolutions but a lot of filters [44]. The Visual Geometry Organization, or VGG for short, was a group that replaced Alex Net which was established in Oxford. It adopts and enhances some concepts from its forerunners and uses deep convolutional neural layers to increase accuracy. Comparatively, managing VGGNet's 138 million parameters can be challenging.

VGG16 has thirteen convolutional layers, five Max Pooling layers, and three Dense layers for a total of twenty-one layers, but only sixteen weight layers or trainable parameters layers [45]. Each of the 16 layers has one convolution and one pooling layer, Fig. 7. VGG16 can be enhanced through transfer learning.

Following the rectified linear unit (ReLU) activations, the image data is transmitted through the first of two convolutional layers with a minimum receiving area of 3X3. In each of these two layers, there are 64 filters. One pixel serves as padding, while one pixel always serves as the convolution step. The first convolutional layer is responsible for capturing low-level information such as gradient and edge orientation, among other information. The spatial maxima are then binned with a step of 2 pixels in a 2x2 pixel window for activation maps. An activation's size is cut in half. Consequently, the activations at the base of the first stack are 112x112x64 long. The activations then proceed via the 128 filters in the second stack as opposed to the 64 in the first one.

The size is 56x56x128 as a result after the second layer. A maximum pool layer and three convolutional layers make up the third layer. Because 256 filters are employed, the output stack size is 28x28x256.

The following two sets of three convolutional layers have each 512 filters. The final stack is of 7x7x512 size for both. Following stacks of convolutional layers with a flattened layer in between are the three fully connected layers. The last completely connected layer serves as the output layer, and has 1000 neurons, or 1000 potential classifications of the ImageNet dataset. The previous two fully connected layers have each 4096 neurons. The SoftMax activation layer, which is utilized for categorization, comes after the output layer. In order to adapt the architecture to high-level characteristics, additional layers are also helpful. The spatial size of the convolved feature is decreased by the pooling layer. The amount of processing power needed to process the data lowers as its dimension increases. Smooth training is made possible by the VGG16 model, which is useful for extracting rotation- and position-invariant dominating features.

2. VGG19

A 19-layer version of the VGG model is known as the VGG19 model, which has 16 convolution layers, three fully connected layers, 5 Max Pool layers, and 1 SoftMax layer, Fig. 8 [46]. An RGB image of fixed size (224*224) was provided to this network as input, indicating that the matrix was of the form (224,224,3). The only preprocessing was to take the mean RGB value for the entire training set and subtract it from each pixel [47]. The complete visual concept was then covered using kernels of size (3*3) with a step size of 1 (one) pixel. Spatial padding was then applied to preserve the spatial resolution of the image. Step two was then used to create maximum pooling in two * 2-pixel windows. Then, instead of using tanh or sigmoid functions, a ReLU was used to induce non-linearity and improve processing speed. Three final connected layers are then implemented, the first two of which are 4096 in size, followed by a 1000-channel ILSVRC classification layer, and finally a SoftMax activation layer, which is used for category classification. It has been used as a good classification architecture for various other datasets. The models were publicly available, so they can be used as is or with minor modifications for other similar work.

3. InceptionV3

Convolutional neural networks are the foundation of the deep learning model known as InceptionV3, which was first developed as a Google network module for image analysis and object detection.

Inception Networks (Google Net/Inception v1) are more cost- and time-effective computationally than VGGNet in terms of the number of network parameters produced. It has 42 layers and a lower error rate than previous models, Fig. 9. To improve model adaptation, the InceptionV3 model uses a number of mesh optimization strategies.

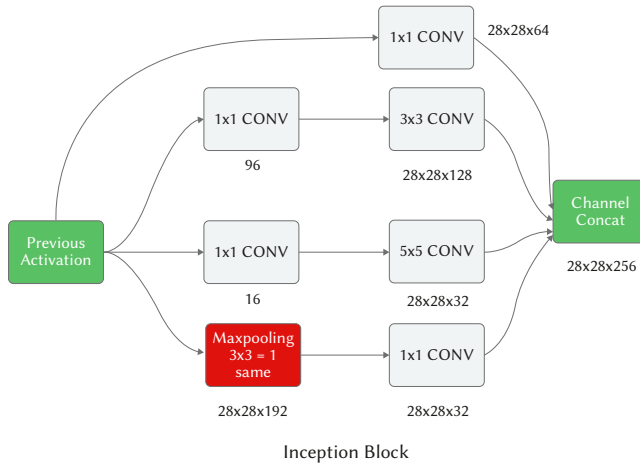


Fig. 9. Layers used in the InceptionV3 model.

The used approaches are factorized convolution, regularization, dimensionality reduction, and parallelized calculations [48]. The number of parameters in the network is decreased via factorized convolutions, which enhances computational effectiveness. It also benefits the network performance. Training becomes faster as smaller convolutions take the place of bigger ones. For instance, replacing a 5 5 convolution with two 3 3 filters only requires 18 (3*3+3*3) parameters. In asymmetric convolutions, a 3 3 convolution can be swapped out for a 1 3 convolution followed by a 3 1 convolution. If the 3 3 convolutions were switched out for a 2 2, there would be a lot more parameters than in the case of the described asymmetric convolution. The network suffers a considerable loss as a result of the losses caused by the little CNN that was added between the layers during training. In InceptionV3, a third classifier acts as a regularization term. Last but not least, pooling procedures are frequently used to achieve a grid size reduction strategy. The final building incorporates all of the principles previously mentioned. The InceptionV3 was used in this study because, while not slower than the Inception V1 and V2 models, it is more effective and has a deeper network [49]. The InceptionV3 model is less expensive to calculate.

In Fig. 10, the proposed customized model is shown. First, the model was built with a standard structure, and later it was fine-tuned for respective models. Methodologies such as feature extraction, data augmentation, and three transfer learning models were used for the comparison purpose in this study. As because of the transfer learning concept is employed therefore there are no overfitting issues with the model. First, input comes from an audio recording of the bird under analysis. After that, features are extracted from the audio input using signal pre-processing techniques. After that, the data augmentation task is accomplished. The processed components are then fed into a powerful classification model and the idea of Transfer Learning to produce the best results for a wide range of species. The used three models are built on pre-trained networks called VGG16, VGG19, and InceptionV3, which were trained using data from 37 different bird species.

In the proposed VGG16 model, there are five convolutions' blocks. Each block contains a convolution 2D model and max-pooling 2D layer. The input of the model is 224, 224 with three dimensions; after one complete convolution, the output size is (112, 112,64). Following another convolution, the output is (56,56,120) after block three, (28,28,256) in partnership four, (14,14,512) in partnership five, and (7,7,512) as input and output are 512 in partnership six. The included dropout layer has a very slight change, and the final dense layer has 256 as an input and 37 as an output. Generally, all the layers of VGG16 were frozen and a customized layer was added. The Sigmoid function is used as an activation function and optimizer. As a loss function, an Adam optimizer with cross-entropy was used. For the VGG19 model, ImageNet was used as a weight, and average pooling, a customized base layer, and convolution layers with 256 dense layers with activation function as ReLu with dropout 0.1 were used. Additionally, SoftMax with a learning rate of 0.00005 was employed in the final layer. During the course of the model-building procedure, the Adam optimizer and the loss function were used as the categorical cross-entropy. Lastly, for the InceptionV3 model, ImageNet was used as a weight, and average pooling was employed. Lastly, a customized model was built by adding custom layers. In the customized model, 256 dense layers with an activation function ReLU were used, and a dropout of 0.4 was used. With the Adam optimizer, the model was built using categorical cross-entropy as the loss function.

IV. EXPERIMENTAL RESULTS & ANALYSIS

We have experimentally chosen three transfer learning models in this study: VGG16, VGG19, and InceptionV3 model. Table II, Table III,

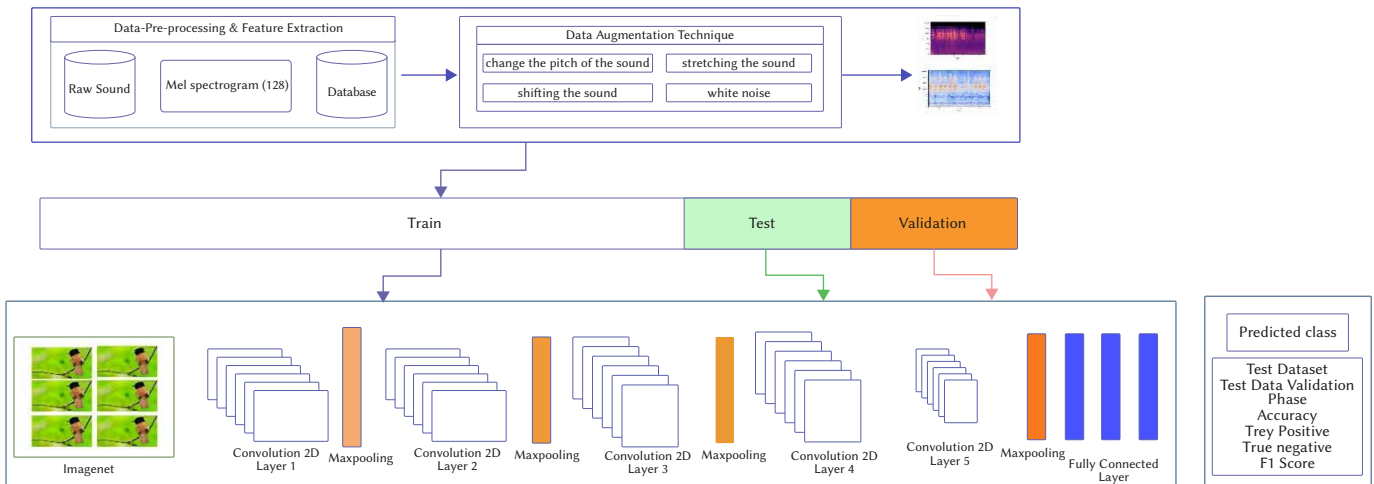


Fig. 10. Proposed deep learning model.

and Table IV show the individual performance of the VGG16, VGG19, and InceptionV3 models respectively. From these three tables, it is observed that the performance of the proposed InceptionV3 model shows better performance when it is compared with the VGG16 and VGG19 models. The experimental results of the InceptionV3 model are reported as 86% precision, 86% of recall, and 85% of F1-score. The classification results of VGG16 are as follows: 18 out of 37 bird sounds: ashy_prinia, brown_fish_owl, brown_winged_kingfisher, cinnamon_bittern, collared_kingfisher, common_woodshrike, fulvous_breasted_woodpecker, grey_headed_fish_eagle, lotens_sunbird, red_whiskered_bulbul, striated_babbler, swamp_francolin, tree_pipit, western_osprey,

asian_openbill, baya_weaver, brown_cheeked_fulvetta, and western_yellow_wagtail, were 100% detected from the test data. The overall accuracy achieved using the VGG16 model was equal to 78%. For the VGG19 model, the training accuracy obtained was of 64%, and the test accuracy was 61.9%. According to the categorization results, VGG19 obtained 100% of recognition in 17 of the 37 test cases. In the InceptionV3 model with a batch size of 32, the obtained train accuracy was 95%, and the test accuracy of 85%. As to the classification results, 24 of the 37 species were 100% detected in the test dataset.

TABLE II. VGG16 CLASSIFICATION MODEL

Class	precision	recall	f1-score	support
0	0.80	0.29	0.42	14
1	1.00	1.00	1.00	12
2	0.73	1.00	0.85	11
3	0.67	0.57	0.62	14
4	1.00	1.00	1.00	14
5	0.50	0.55	0.52	11
6	0.42	0.62	0.50	13
7	0.71	0.62	0.67	16
8	1.00	1.00	1.00	12
9	1.00	1.00	1.00	10
10	1.00	1.00	1.00	10
11	0.93	1.00	0.96	13
12	0.87	1.00	0.93	13
13	0.80	0.67	0.73	18
14	0.90	0.75	0.82	12
15	1.00	0.85	0.92	13
16	1.00	1.00	1.00	13
17	1.00	1.00	1.00	11
18	0.50	0.25	0.33	12
19	0.70	0.54	0.61	13
20	0.62	0.45	0.53	11
21	0.56	0.42	0.48	12
22	1.00	1.00	1.00	11
23	0.37	0.64	0.47	11
24	0.92	1.00	0.96	12
25	1.00	0.46	0.63	13
26	0.92	0.92	0.92	12
27	0.50	0.42	0.45	12
28	0.37	0.64	0.47	11
29	1.00	1.00	1.00	10
30	0.45	0.77	0.57	13
31	0.71	1.00	0.83	12
32	1.00	1.00	1.00	12
33	1.00	1.00	1.00	11
34	1.00	1.00	1.00	10
35	0.92	0.85	0.88	13
36	1.00	1.00	1.00	10
accuracy			0.78	451
macro avg	0.81	0.79	0.78	451
weighted avg	0.80	0.78	0.78	451

TABLE III. VGG19 CLASSIFICATION MODEL

Class	precision	recall	f1-score	support
0	0.33	0.36	0.34	14
1	1.00	1.00	1.00	12
2	1.00	0.36	0.53	11
3	0.26	0.36	0.30	14
4	1.00	1.00	1.00	14
5	0.50	0.18	0.27	11
6	0.67	0.15	0.25	13
7	0.39	0.44	0.41	16
8	0.60	1.00	0.75	12
9	0.91	1.00	0.95	10
10	1.00	1.00	1.00	10
11	0.62	1.00	0.76	13
12	0.93	1.00	0.96	13
13	0.67	0.22	0.33	18
14	0.33	0.67	0.44	12
15	0.86	0.46	0.60	13
16	0.92	0.85	0.88	13
17	1.00	1.00	1.00	11
18	0.40	0.17	0.24	12
19	0.50	0.31	0.38	13
20	0.44	0.64	0.52	11
21	0.52	0.92	0.67	12
22	0.91	0.91	0.91	11
23	0.33	0.36	0.35	11
24	0.57	1.00	0.73	12
25	0.15	0.15	0.15	13
26	0.67	0.83	0.74	12
27	0.43	0.25	0.32	12
28	0.00	0.00	0.00	11
29	0.77	1.00	0.87	10
30	1.00	0.23	0.38	13
31	0.50	0.58	0.54	12
32	0.71	1.00	0.83	12
33	0.86	0.55	0.67	11
34	1.00	1.00	1.00	10
35	0.50	0.54	0.52	13
36	0.53	1.00	0.69	10
accuracy			0.62	451
macro avg	0.64	0.63	0.6	451
weighted avg	0.64	0.62	0.59	451

A. Accuracy

In order to meaningfully evaluate a machine learning model's performance, accuracy is a metric frequently used. A model's accuracy is usually calculated once the parameters are specified and represented in terms of percentage, which is a statistic that shows how accurately the model's performance contrasts with actual data. Figs. 11, 12, and 13 show the accuracy curves for the built learning models. In the experiment using the baseline model of VGG16, only ten epochs with a batch size 32 were run, and the obtained train accuracy was of 75% and the test one was 78%. Similarly, the VGG19 model also run in 10 periods with a batch size of 32, and 64% was train accuracy and 61.9% was test accuracy and 61.9%

TABLE IV. INCEPTIONV3 CLASSIFICATION MODEL

Class	precision	recall	f1-score	support
0	0.46	0.43	0.44	14
1	1.00	1.00	1.00	12
2	0.85	1.00	0.92	11
3	0.88	1.00	0.93	14
4	0.93	1.00	0.97	14
5	0.35	0.55	0.43	11
6	0.90	0.69	0.78	13
7	0.70	0.88	0.78	16
8	1.00	1.00	1.00	12
9	1.00	1.00	1.00	10
10	1.00	1.00	1.00	10
11	1.00	1.00	1.00	13
12	1.00	1.00	1.00	13
13	0.91	0.56	0.69	18
14	1.00	1.00	1.00	12
15	1.00	1.00	1.00	13
16	0.93	1.00	0.96	13
17	0.92	1.00	0.96	11
18	0.75	0.50	0.60	12
19	0.78	0.54	0.64	13
20	0.92	1.00	0.96	11
21	0.86	1.00	0.92	12
22	0.79	1.00	0.88	11
23	0.83	0.91	0.87	11
24	1.00	1.00	1.00	12
25	0.71	0.77	0.74	13
26	1.00	1.00	1.00	12
27	0.43	0.25	0.32	12
28	0.30	0.27	0.29	11
29	0.91	1.00	0.95	10
30	0.90	0.69	0.78	13
31	1.00	1.00	1.00	12
32	1.00	1.00	1.00	12
33	0.92	1.00	0.96	11
34	1.00	1.00	1.00	10
35	1.00	1.00	1.00	13
36	0.91	1.00	0.95	10
accuracy			0.86	451
macro avg	0.86	0.87	0.86	451
weighted avg	0.86	0.86	0.85	451

the test one. On the InceptionV3 model, which was used in further experiments, with 10 epochs, a training accuracy of 95% and a test accuracy of 85% were obtained.

B. Loss

A more accurate model is indicated by lower loss values. The loss is not expressed as a percentage, in contrast to accuracy. The built learning models' loss curves are shown in Figs. 14, 15 and 16. The training loss of the VGG16, VGG19, and InceptionV3 models decreased over time, but the validation data revealed frequent variations and substantial loss. The loss function shown was in the 0.9 to 1.5 in range in the three studied models. In the training of the studied models, categorical_crossentropy was used.

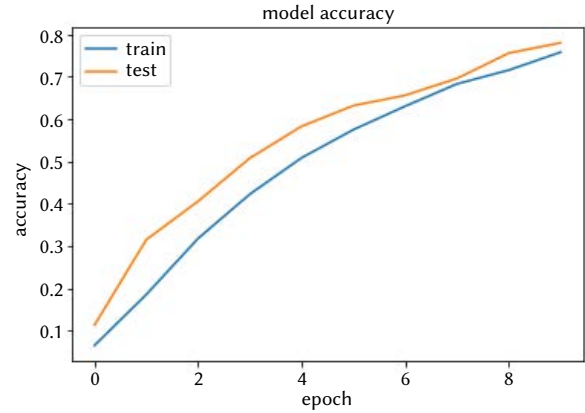


Fig. 11. VGG16 model's accuracy.

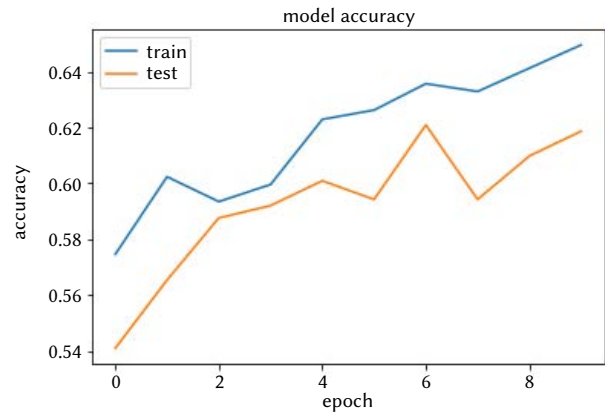


Fig. 12. VGG19 model's accuracy.

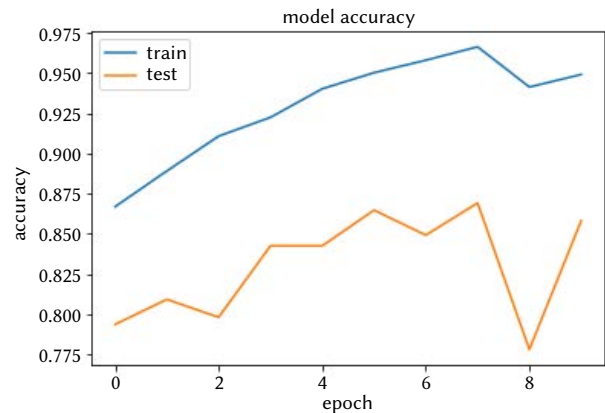


Fig. 13. INCEPTIONV3 model's accuracy.

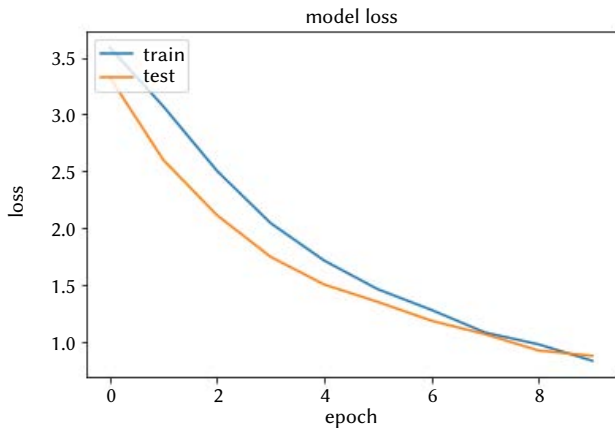


Fig. 14. VGG16 model's loss.

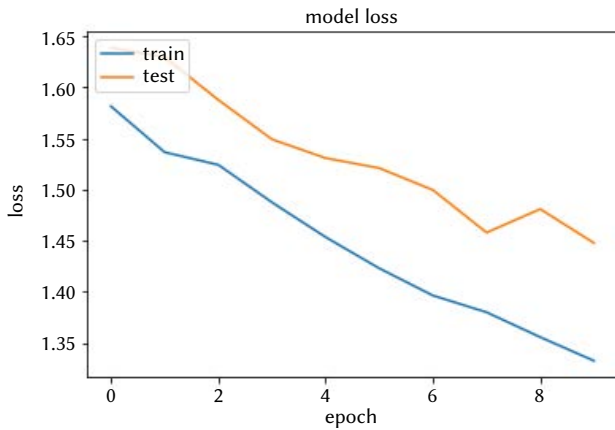


Fig. 15. VGG19 model's loss.

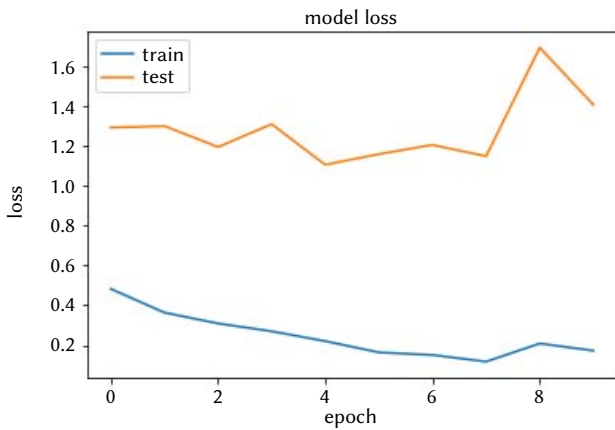


Fig. 16. INCEPTIONV3 model's loss.

C. Confusion Matrix

An evaluation of the performance of a classification model, or “classifier”, on a set of test data for which the true values are known is given by a confusion matrix, which is a table. The matrix also allows a comparison between the targets’ actual values and the model projections. To properly comprehend the classification findings, the confusion matrix for each of the three classification architectures were built, Figs. 17, 18 and 19.

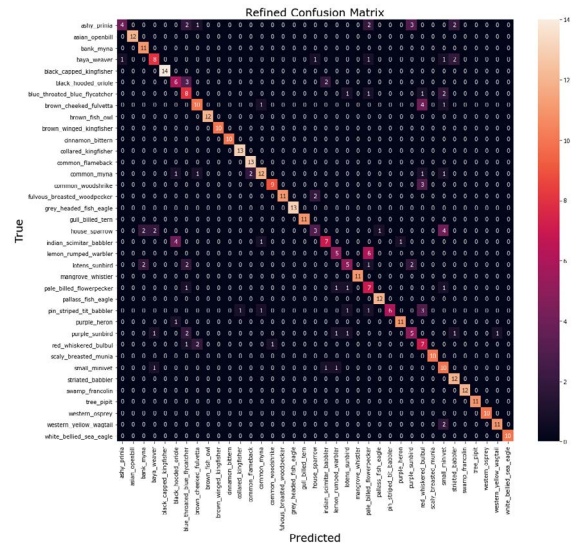


Fig. 17. Confusion matrix obtained by VGG16.

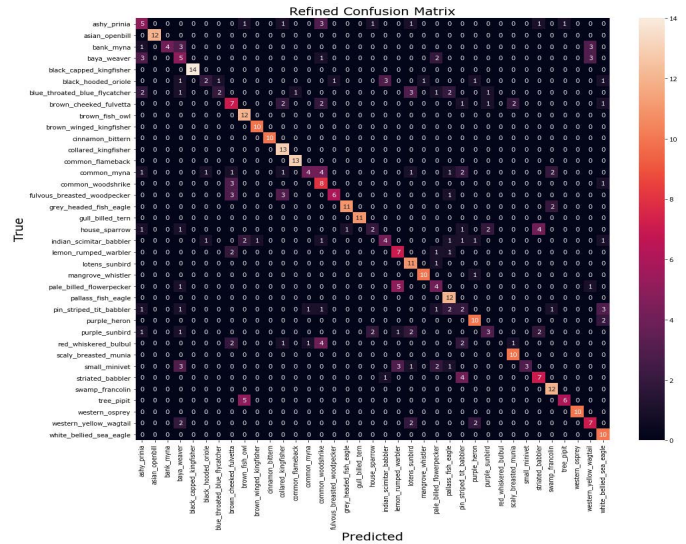


Fig. 18. Confusion matrix obtained by VGG19.

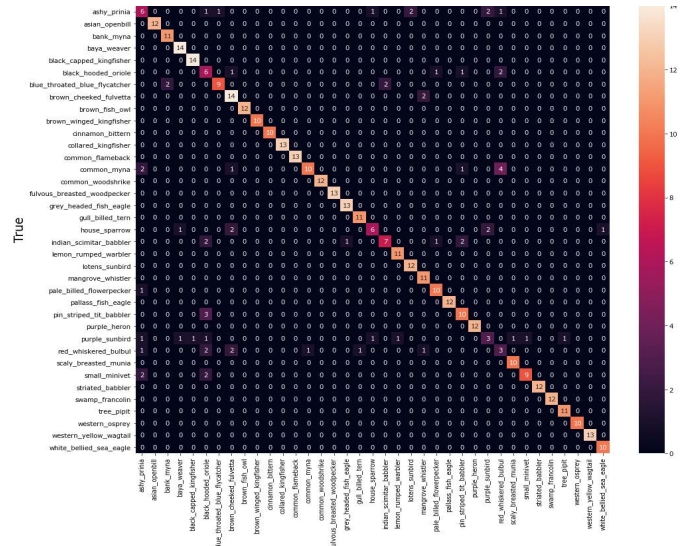


Fig. 19. Confusion matrix obtained by INCEPTIONV3.

V. DISCUSSION

Working with bird species of a more significant number of types is challenging. Deep learning architectures have improved speech recognition accuracy, and automated learning approaches have been developed. Transfer learning technique was used in this study as 37 bird species were addressed, and a large dataset with a wide range of bird sounds is required. In this research, the categorization of bird noises is accomplished via the utilization of three different deep-learning frameworks. These frameworks are VGG16, VGG19, and InceptionV3. All the models use the same model parameters. As was shown, the InceptionV3 model obtained the best result. However, M Lasseck et al. [50] showed 93% accuracy using ensemble models with deep convolution neuronal networks with a pre-trained model but using a more significant number of epochs. The proposed model outperforms the solutions proposed by earlier work that was carried out by other researchers. Previously, various models gave a visual representation of the sound, but the suggested model is capable of working directly with the unprocessed audio file. According to another finding, the InceptionV3 model performs better than the other two models in this regard. In addition, acoustic properties were gathered from bird calls and were classified using various feature extraction techniques. It has been demonstrated that the proposed strategy is capable of boosting prediction accuracy. A novel method for identifying a large number of bird species in the Sundarban region of West Bengal, India was devised using existing recordings of their sounds.

VI. CONCLUSION

The suggested model may be put into low-cost devices via the use of a technique that is both cost-effective and scalable; hence, more devices can be employed to cover more land. In this experiment, it was shown that a transfer-learned network that had previously been trained on ImageNet shows a better predictive capability and accelerates convergence when compared with the same network architecture that is trained from scratch. The experiment was conducted in order to demonstrate this. When there are a limited number of high-quality datasets available, it is advantageous to utilize a model that has already been pre-trained because of the benefits it provides. In terms of practical uses, the suggested approach may be of great use to ornithologists by making the identification of bird species a straightforward process. In the future, in order to enhance the recognition rate, we want to use a variety of noise reduction filtering techniques during the pre-processing step. In addition, another problem that should be addressed is the overlapping of sounds.

REFERENCES

- [1] M. A. Tabur and Y. Ayvaz, "Ecological importance of birds," in Second International Symposium on Sustainable Development Conference, 2010, Jun., pp. 560-565.
- [2] S. D. H. Permana et al., "Classification of bird sounds as an early warning method of forest fires using Convolutional Neural Network (CNN) algorithm," *Journal of King Saud University - Computer and Information Sciences*, Inf. Sci., 2021.
- [3] G. F. Budney and R. W. Grotke, "Techniques for audio recording vocalizations of tropical birds," *Ornithological Monographs*, no. 48, pp. 147-163, 1997, doi:10.2307/40157532.
- [4] Available at: <https://www.environmentalscience.org/birds-environmental-indicators> (last access date: 18/18/2022).
- [5] Available at: <https://www.ck12.org/biology/bird-ecology/lesson/Importance-of-Birds-MS-LS/> (last access date: 18/12/2022).
- [6] Available at: <https://www.thespruce.com/bird-courtship-behavior-386714> (last access date: 18/12/2022).
- [7] Available at: <https://www.birdlife.org/worldwide/news/why-we-need-birds-far-more-they-need-us> (last access date: 18/12/2022).
- [8] S. Fagerlund, "Bird species recognition using support vector machines," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 1-8, 2007, doi:10.1155/2007/38637.
- [9] N. Das et al., *Machine Learning Models for Bird Species Recognition Based on Vocalization: A Succinct Review*. Information Technology and Intelligent Transportation Systems, 2020, pp. 117-124.
- [10] C. Yüksel, 2020, Bird call detection using deep learning (Master's thesis, Fen Bilimleri Enstitüsü).
- [11] S. Bhattacharya et al., "Deep classification of sound: A concise" in *Proceeding of First Doctoral Symposium on Natural Computing Research: DSNCR 2020*, vol. 169. Springer Nature, 2021, Mar.
- [12] N. Das et al., "Building of an edge-enabled drone network ecosystem for bird species identification," *Ecological Informatics*, vol. 68, p. 101540, 2022, doi: 10.1016/j.ecoinf.2021.101540.
- [13] S. Bhattacharya et al., "Deep analysis for speech emotion recognition" in *Second International Conference on Computer Science, Engineering and Applications (ICCSEA)*, vol. 2022. IEEE, 2022, Sept., pp. 1-6, doi:10.1109/ICCSEA54677.2022.9936080.
- [14] K. Lan et al., "A survey of data mining and deep learning in bioinformatics," *Journal of Medical Systems*, vol. 42, no. 8, pp. 139, 2018, doi:10.1007/s10916-018-1003-9.
- [15] Y. Wu et al., "Learning models for semantic classification of insufficient plantar pressure images," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 1, pp. 51-61, 2020, doi:10.9781/ijimai.2020.02.005.
- [16] H. Chang et al., "Unsupervised transfer learning via multi-scale convolutional sparse coding for biomedical applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1182-1194, 2018, doi:10.1109/TPAMI.2017.2656884.
- [17] R. Wald et al., "Hidden dependencies between class imbalance and difficulty of learning for bioinformatics datasets" in *14th International Conference on Information Reuse & Integration (IRI)*, vol. 2013. IEEE, 2013, Aug., pp. 232-238, doi:10.1109/IRI.2013.6642477.
- [18] C. Tan et al., "A survey on deep transfer learning" in *International conference on artificial neural networks*. Cham: Springer, 2018, Oct., pp. 270-279.
- [19] L. Muda et al., 2010, Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. arXiv preprint arXiv:1003.4083.
- [20] C. Y. Koh et al., 2019, Sept., "Bird sound classification using convolutional neural networks" in *Clef [Working notes]*.
- [21] A. Fritzier et al., 2017, "Recognizing bird species in audio files using transfer learning" in *Clef [Working notes]*.
- [22] S. Ntalampiras, "Bird species identification via transfer learning from music genres," *Ecological Informatics*, vol. 44, pp. 76-81, 2018, doi: 10.1016/j.ecoinf.2018.01.006.
- [23] D. B. Efreanova et al., "Data-efficient classification of birdcall through convolutional neural networks transfer learning" in *Digital Image Computing: Techniques and Applications (DICTA)*, vol. 2019. IEEE, 2019, Dec., pp. 1-8, doi:10.1109/DICTA47822.2019.8946016.
- [24] J. Bai et al., 2019, "Inception-v3 based method of LifeCLEF," vol. 2019 Bird Recognition in Clef [Working notes].
- [25] M. Zhong et al., "Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling," *Applied Acoustics*, vol. 166, p. 107375, 2020, doi: 10.1016/j.apacoust.2020.107375.
- [26] A. K. Ibrahim et al., "Transfer learning for efficient classification of grouper sound," *Journal of the Acoustical Society of America*, vol. 148, no. 3, pp. EL260, 2020, doi:10.1121/1.510001943.
- [27] R. Rajan and A. Noumida, "Multi-label bird species classification using transfer learning" in *2021 International Conference on Communication, Control and Information Sciences (ICCISc)*, vol. 1. IEEE, 2021, Jun., doi:10.1109/ICCISc52257.2021.9484858.
- [28] E. J. Henri and Z. Mungloo-Dilmohamad, "A deep transfer learning model for the identification of bird songs: A case study for Mauritius" in *International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, vol. 2021. IEEE, 2021, Oct., pp. 1-6, doi:10.1109/ICECCME52200.2021.9590917.
- [29] K. W. Gunawan et al., "A transfer learning strategy for owl sound

classification by using image classification model with audio spectrogram,” International Journal on Electrical Engineering and Informatics, vol. 13, no. 3, pp. 546-553, 2021, doi:10.15676/ijeel.2021.13.3.3.

- [30] S. Nayak et al., “Whose hoot? Identification of owl species using call recognition with neural networks,” SSRN Journal, 2022, doi:10.2139/ssrn.4020038.
- [31] N. Sharma et al., “Automatic identification of bird species using audio/video processing” in International Conference for Advancement in Technology (ICONAT), vol. 2022. IEEE, 2022, Jan., pp. 1-6, doi:10.1109/ICONAT53423.2022.9725906.
- [32] Y. Kumar et al., “A novel deep transfer learning models for recognition of birds sounds in different environment,” Soft Computing, pp. 1-14, 2022.
- [33] S. Bhattacharya et al., “Emotion detection from multilingual audio using deep analysis,” Multimedia Tools and Applications, pp. 1-30, 2022.
- [34] E. Sprengel et al., 2016, Audio-based bird species identification using deep learning techniques (No. CONF, pp. 547-559).
- [35] E. Cakir et al., “Convolutional recurrent neural networks for bird audio detection,” 25th European Signal Processing Conference EUSIPCO, vol. 2017, 2017. 2017-Janua, pp. 1744-1748, doi:10.23919/EUSIPCO.2017.8081508.
- [36] J. Kim et al., “Acoustic classification of mosquitoes using convolutional neural networks combined with activity circadian rhythm information,” International Journal of Interactive Multimedia and Artificial Intelligence, vol. 7, no. 2, 2021, doi:10.9781/ijimai.2021.08.009.
- [37] S. Ahuja et al., “Deep transfer learning-based automated detection of COVID-19 from lung CT scan slices,” Applied intelligence (Dordrecht, Netherlands), vol. 51, no. 1, pp. 571-585, 2021, doi:10.1007/s10489-020-01826-w.
- [38] M. Singh et al., “Transfer learning-based ensemble support vector machine model for automated COVID-19 detection using lung computerized tomography scan data,” Medical & Biological Engineering & Computing, vol. 59, no. 4, pp. 825-839, 2021, doi:10.1007/s11517-020-02299-2.
- [39] D. A. Pitaloka et al., “Enhancing CNN with preprocessing stage in automatic emotion recognition,” Procedia Computer Science, vol. 116, pp. 523-529, 2017 [doi:10.1016/j.procs.2017.10.038].
- [40] M. Morrison et al., 2021, Neural pitch-shifting and time-stretching with controllable LPCNet. arXiv preprint arXiv:2110.02360.
- [41] P. B. Baptista and C. Antunes, “Bioacoustic classification framework using transfer learning,” Model Decision Artificial Intelligence, vol. 35, 2021.
- [42] A. Bhaik et al., “Detection of improperly worn face masks using deep learning-A preventive measure against the spread of COVID-19,” International Journal of Interactive Multimedia and Artificial Intelligence, vol. 7, no. 7, 2021, doi:10.9781/ijimai.2021.09.003.
- [43] K. He et al., “Deep residual learning for image recognition” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2016, vol. 7, no. 3, 2016, pp. 770-778, doi:10.1109/CVPR.2016.90
- [44] Available at: <https://iq.opengenus.org/vgg16/> [Last Access Date: 18.12.2022].
- [45] S. K. Rahut et al., “Bengali abusive speech classification: A transfer learning approach using” VGG-16 in Emerging Technology in Computing, Communication and Electronics (ETCCE), vol. 2020. IEEE, 2020, Dec., pp. 1-6.
- [46] A. Ashurov et al., “Environmental sound classification based on transfer-learning techniques with multiple optimizers,” Electronics, vol. 11, no. 15, p. 2279, 2022 [doi:10.3390/electronics11152279].
- [47] M. J. Horry et al., “COVID-19 detection through transfer learning using multimodal imaging data,” IEEE Access, vol. 8, pp. 149808-149824, 2020 [doi:10.1109/ACCESS.2020.3016780].
- [48] Available at: <https://blog.paperspace.com/popular-deep-learning-architectures-resnet-inceptionv3-squeezenet/> [Last Access Date: 18.12.2022].
- [49] Y. Shen et al., “Urban acoustic classification based on deep feature transfer learning,” Journal of the Franklin Institute, vol. 357, no. 1, pp. 667-686, 2020 [doi:10.1016/j.jfranklin.2019.10.014].
- [50] M. Lasseck, “Acoustic bird detection with deep convolutional neural networks” in Proc. Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), 2018, Nov., pp. 143-147.



Nabanita Das

Nabanita Das is a Ph.D. Research Scholar with the Department of Computer science and engineering, GIET University, Gunupur, Orissa, India. Currently, she is an Asst. Professor in the Department of Computer Science and Engineering, Bengal Institute of Technology, India. She received the M. Tech. degree from MAKAUT, West Bengal, India, and has more than ten years of teaching experience. She is actively involved in research in the domains of Machine Learning, Deep Learning, IoT, Software Engineering, and Computer Aided Diagnosis.



Neelamadhab Padhy

Neelamadhab Padhy received his Ph.D. in 2018 from Sri Satya Sai University of technology and medical science, Sehore, India. He is now employed as an Associate Professor in the Department of Computer science and engineering, GIET University, Gunupur. His research topics are machine learning, deep learning software engineering, image processing, etc. He published more than 30 peer-reviewed journal and conference papers. He is a life member of CSI and a member of the IE and Soft Computing Society.



Nilanjan Dey

Nilanjan Dey is an Associate Professor in the Department of Computer Science and Engineering, Techno International New Town, Kolkata, India. He is a visiting fellow of the University of Reading, UK. He also holds a position of Adjunct Professor at Ton Duc Thang University, Ho Chi Minh City, Vietnam. Previously, he held an honorary position of Visiting Scientist at Global Biomedical Technologies Inc., CA, USA (2012–2015). He was awarded his PhD from Jadavpur University in 2015. He is the Editor-in-Chief of the International Journal of Ambient Computing and Intelligence, IGI Global, USA. He is the Series Co-Editor of Springer Tracts in Nature-Inspired Computing (SpringerNature), Data-Intensive Research (SpringerNature), Advances in Ubiquitous Sensing Applications for Healthcare (Elsevier). He is an associate editor of IET Image Processing and editorial board member of Complex & Intelligent Systems, Springer Nature, Applied Soft Computing, Elsevier etc. He is working in the area of medical imaging, machine learning, computer aided diagnosis, data mining, etc. He is the Indian Ambassador of the International Federation for Information Processing—Young ICT Group and Senior member of IEEE.



Sudipta Bhattacharya

Sudipta Bhattacharya is an Asst. Professor in the Department of Computer Science and Engineering, Bengal Institute of Technology, India. He is a Ph.D. Research Scholar with the Department of Computer science and engineering, GIET University, Gunupur, Orissa, India. He received his Bachelor of Technology, (IT) from West Bengal University of Technology, India, and Master of Technology (IT) from the Indian Institute of Engineering Science and Technology, Shibpur, India. His area of research interest is Pattern Recognition and Speech emotion Recognition.



João Manuel R.S. Tavares

João Manuel R.S. Tavares received the degree in mechanical engineering and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the Universidade do Porto, Portugal, in 1992, 1995, and 2001, respectively, and the Habilitation degree in mechanical engineering, in 2015. He is currently a Senior Researcher in the Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial (INEGI), and a Full Professor in the Department of Mechanical Engineering (DEMec), Faculdade de Engenharia da Universidade do Porto (FEUP). He is the co-editor of more than 80 books, the co-author of more than 50 book chapters, 650 articles in international and national journals and conferences, and three international and three national patents. He has been a committee member of several international and national journals and conferences. He is the Co-Founder and the Co-Editor of the book series Lecture Notes in Computational Vision and Biomechanics (Springer), the Founder and Editor-

in-Chief of the journal *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* (Taylor & Francis), the Editor-in-Chief of the journal *Computer Methods in Biomechanics and Biomedical Engineering* (Taylor & Francis), and the Co-Founder and the Co-Chair of the International Conference Series, such as *CompIMAGE*, *ECCOMAS VipIMAGE*, *ICCEBS*, and *BioDental*. Additionally, he has been the co-supervisor of several M.Sc. and Ph.D. thesis and a supervisor of several postdoctoral projects. He has participated in many scientific projects both as a Researcher and as a Scientific Coordinator. His research interests include computational vision, medical imaging, computational mechanics, scientific visualization, human-computer interaction, and new product development. (More information can be found at <https://www.fe.up.pt/~tavares>).

Deobfuscating *Leetspeak* With Deep Learning to Improve Spam Filtering

Iñaki Vélez de Mendizabal¹, Xabier Vidriales¹, Vitor Basto-Fernandes², Enaitz Ezpeleta¹, José R. Méndez^{3,4,5,*}, Urko Zurutuza¹

¹ Mondragon Unibersitatea, Faculty of Engineering, Electronics and Computing Department. Loramendi 4, Modragon 20500 Gipuzkoa (Spain)

² Instituto Universitário de Lisboa (ISCTE-IUL), University Institute of Lisbon, ISTAR-IUL, Av. das Forças Armadas, 1649-026 Lisboa (Portugal)

³ Department of Computer Science, ESEI - Escola Superior de Enxeñaría Informática, Universidade de Vigo, Campus Universitario As Lagoas s/n, 32004 Ourense (Spain)

⁴ CINBIO - Biomedical Research Centre, Universidade de Vigo, Campus Universitario Lagoas-Marcosende, 36310-Vigo, Pontevedra (Spain)

⁵ SING Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur), Hospital Álvaro Cunqueiro Bloque técnico, Estrada de Clara Campoamor, 36312-Vigo, Pontevedra (Spain)

Received 31 March 2022 | Accepted 30 May 2023 | Published 7 July 2023



ABSTRACT

The evolution of anti-spam filters has forced spammers to make greater efforts to bypass filters in order to distribute content over networks. The distribution of content encoded in images or the use of *Leetspeak* are concrete and clear examples of techniques currently used to bypass filters. Despite the importance of dealing with these problems, the number of studies to solve them is quite small, and the reported performance is very limited. This study reviews the work done so far (very rudimentary) for *Leetspeak* deobfuscation and proposes a new technique based on using neural networks for decoding purposes. In addition, we distribute an image database specifically created for training *Leetspeak* decoding models. We have also created and made available four different corpora to analyse the performance of *Leetspeak* decoding schemes. Using these corpora, we have experimentally evaluated our neural network approach for decoding *Leetspeak*. The results obtained have shown the usefulness of the proposed model for addressing the deobfuscation of *Leetspeak* character sequences.

KEYWORDS

Convolutional Neural Networks, Deep Learning, *Leetspeak*, Spam Filtering, Text Deobfuscation.

DOI: 10.9781/ijimai.2023.07.003

I. INTRODUCTION

CURRENTLY, the Internet is one of the most widely used means of communication for exchanging personal (e.g. recreational activities) and corporate information (e.g. business topics). In July 2020, there were more than 4.57 billion Internet users, of which almost 4 billion were using social media services (<https://www.statista.com/statistics/617136/digital-population-worldwide/>). Internet users can enjoy the speed and simplicity of exchanging information, shopping online or contacting other users. However, some users use the Internet unethically for their benefit, degrading the experience of other users. In particular, one of the most annoying abuses is the distribution of inappropriate and unsolicited content (spam) through communication services based on the exchange of text messages such as classic email [1], [2], social networks [1], [3] or instant messaging [4], [5].

The growth of spam on the Internet has generated the need to develop sophisticated text classification techniques that must be highly

reliable and fast to operate. They are used to automatically classify messages into two different spam and ham (legitimate) categories by combining information retrieval [6] (IR) and Machine Learning (ML) [7]. Many text classification approaches have been widely applied to address the problem. Some initial approaches exploited Bag of Words (BoW) representation schemes (using frequency, binary or inverse document frequency values) in conjunction with different types of classifiers, including (i) Naïve Bayes [8], [9], (ii) memory based approaches [10], (iii) decision trees [11], [12], Support Vector Machines (SVM) [13], Artificial Neural Networks (ANN) [14], logistic regression [15], Artificial Immune Systems (AIS) [16], Boosting of trees [17] and other hybrid methods. The latest advances to improve the performance of this type of classifiers consist of the use of synsets obtained from ontological dictionaries such as Wordnet [18] and Babelnet [19] and different types of semantic processing of words [20]–[22].

In the context of the fight against spam, spammers introduced a lot of tricks to avoid spam filters. One of the best known was the use of attached images which became popular in 2007 [23]. This method relates to attaching images that cannot be processed by text classifiers; but are human understandable spam texts. Fig. 1 shows images with embedded texts that are clearly spam and will not be analysed by text filters.

E-mail address: moncho.mendez@uvigo.es

To combat this type of spam, some researchers took advantage of Optical Character [24] Recognition (OCR) which was initially effective in identifying some words in the image. Later, Battista *et al.* [25] showed how to evade anti-spam filters using text obfuscation techniques in the attached images. To increase the difficulty of identifying the text embedded in the images, spammers add noise to the image [26] (see right image in Fig. 1). More recently, new image-based obfuscation tricks were developed (e.g. as CAPTCHA -Completely Automated Public Turing test to tell Computers and Humans Apart- [27], which can display text and make it unreadable for automatic text recognition systems). However, the latest advances in ANNs have allowed the recognition of the texts [28], [29] included in these CAPTCHAs.

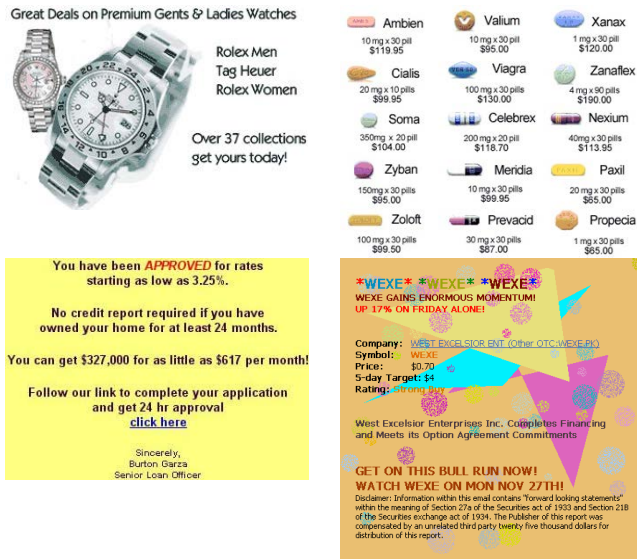


Fig. 1. Examples of images attached to spam messages. These images are part of Image Spam Dataset (https://www.cs.jhu.edu/~mdredze/datasets/image_spam/).

Another important challenge in spam filtering is the recognition of *Leetspeak* (also known as *leet*, *leet text* or *1337*). This type of slang writing has been used since 1980 [30] and consists of replacing some characters with visually similar symbols so that the reader can understand the message. This type of encoding achieves two simultaneous effects: (i) it prevents the classifier from recognising, tokenizing and processing the word and (ii) it produces a Bayesian Poisoning [31] attack that inserts random and apparently harmless words into spam messages, causing a spam email to be incorrectly classified as ham. Table I presents twelve *Leetspeak* representations for the word “viagra” (which is often included in spam messages) out of the approximately 600 trillion possible forms for this word. Each column in Table I shows possible replacements for a single character in the word.

TABLE I. EXAMPLES OF *LEETSPeAK* FORMS FOR THE WORD “VIAGRA”

Original Word	Transformation examples
viagra	V\iagra, /iagra
viagra	v1agra, v;agra
viagra	vi4gra, vi/\gra
viagra	via6ra, via(_-ra
viagra	viag12a, viag/2a
viagra	viagr/, viagr/-\

Table I shows, *Leetspeak* exploits punctuation marks or symbols to hide characters. The replacements made cause misrecognition and misrepresentation of the word during the classification process; therefore, the spammer can bypass spam filters. Using *Leetspeak*, any character (e.g. “A”) can be encoded in many ways and using a different number of symbols (“/-\”, “4”, “|-\”, ...). As *Leetspeak* does not consist of a limited set of symbols, it cannot be solved using a dictionary.

Some previous studies have addressed this problem. Tundis *et al.* designed a convolutional neural network (CNN) to directly classify texts using *Leetspeak* encoding [32]. The use of direct text classification strategies has limitations since the response obtained is not justified. Instead of directly classifying the text, it would be desirable for CNN to allow decoding of the hidden characters in order to provide a solution more understandable from a human point of view. These types of solutions are included in explainable artificial intelligence (XAI) [33]. Subsequently, the same authors proposed a new algorithm for the classification of obfuscated texts that meet the principles of XAI [34]. To do so, they designed a rule-based algorithm in which they exploit a low-precision CNN (*rule-2*) that was created using Chars74K [35] image dataset and a collection of images representing non-english characters. Authors tested their CNN using a train-test experiment with their dataset (Chars74K + non-english chars) achieving a performance up to 94,3 percent. However, the performance of their CNN is not measured by classifying *Leetspeak* sequences. In the context of this study, we have trained different CNN models to identify obfuscated characters using the Chars74K dataset for training. All these models returned low *accuracy* scores (in the interval of 42%-52%). The combination of strategies (rules) seems to have allowed the authors to improve the quality of the results obtained. Taking into account the advances achieved in the context of Deep Learning (DL) applied to solving similar problems [36]–[38], we believe that we can obtain performance improvements by creating CNN models from better training data.

In this study, we are introducing a new computer vision approach based exclusively on the use of a CNN model [39], [40] to decode *Leetspeak*. It is able to accurately identify sequences of *Leetspeak* encoded characters represented as images. Using this approach, we are able to recognize the obfuscated words and thus make the full text available to the spam filter. For the implementation, we used TensorFlow [41] and Keras [42]. Our contributions are: (i) an image database used for training CNNs for *Leetspeak* deobfuscation, (ii) an empirical demonstration that *Leetspeak* recognition can be accurately performed using only CNNs and (iii) datasets for the evaluation of *Leetspeak* decoding schemes.

The rest of the manuscript is structured as follows: Sections II and III describe the materials and methods used to complete this study; Section IV presents and analyses the experimental results and finally, Section V describes the conclusions and future research directions.

II. MATERIALS

Currently, there is no dataset available that contains text messages with words obfuscated using *Leetspeak*. In order to create DL models to decode *Leetspeak* character sequences, we had to create a large set of character images to train neural networks (create models) for the task of recognizing the obfuscated characters. The process of creating the image database is described in first subsection. Additionally, we had to generate a new dataset containing obfuscated messages that can be used as a basis for experimentation on this problem. The second subsection describes the process followed to obtain a dataset containing obfuscated messages using *LeetSpeak*.

¹ Available at <http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/>

A. Training Image Database

This paper introduces a computer vision system based on the use of DL to identify obfuscated characters. The process of creating models capable of decoding *Leetspeak* requires the existence of a set of labelled images in which characters are represented. Following the results of the study conducted by Tundis *et al.* [34], we evaluated the Chars74K image dataset. However, this dataset is oriented to assist in the character recognition in natural images and does not fully fit the target of our study. To validate this statement, we trained some models using the Chars74K dataset and applied them for *Leetspeak* deobfuscation achieving classification accuracies in the interval of 42%-52%. Therefore, we have created a database of character images that will be used to train more efficient models.

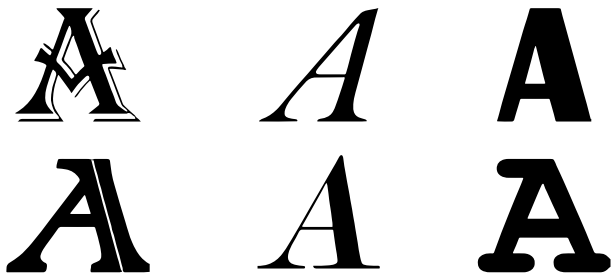


Fig. 2. Examples of images labelled with 'A' character.

Our image database was generated by representing each character ('A'-'Z') using 158 different computer fonts and regular, italic, bold and italic+bold styles. The images were obtained at a resolution of 100x100 pixels. Fig. 2 shows some of the images included in the database and labelled with the character "A".

We improved our image database by adding images extracted from an English handwriting *Dataset*. The resulting image database has a balanced number of images. For each of the 26 characters ('A'-'Z') we obtained at least 632 different images and up to 767.

This set of images has been made available in the community section of Mondragon Unibertsitatea website and in Zenodo [43].

B. Datasets for Evaluating Text Deobfuscation Methods

This subsection describes the method designed to obtain corpora in which spam texts may contain obfuscated words and thus be suitable for evaluating the performance of new deobfuscation processes. To do so, we take advantage of well-known and publicly available spam corpora. Table II compiles a set of well-known corpora that provides some interesting features such as content description, ham/spam ratio and the Universal Resource Locator (URL) where the dataset is available.

As shown in Table II, a large collection of datasets with different sizes and contents are available to test the performance. We selected two datasets with classified YouTube comments (YouTube Comments Dataset and YouTube Spam Collection Dataset) that we had used in a previous study [44]. In this study, we only used a subset of 4000 comments from YouTube Comments Dataset (1000 spam and 3000 ham) while the YouTube Spam Collection Dataset was fully used. As the same datasets are used in both studies, it is possible to compare the results obtained. In addition, to extend the study to the email domain, we also selected two medium-sized email datasets (CSDMC 2010 Spam Corpus and TREC 2007 Public Corpus). In this study, the CSDMC 2010 dataset is fully used, while 4327 (32% of them spam) messages were randomly selected from the TREC 2007 dataset. Therefore, both email corpora have the same length and ham/spam ratio.

Once base datasets were selected, we designed an algorithm to create obfuscated contents to be used for the evaluation of our proposal. Table III exemplifies some replacements used in *Leetspeak* to obfuscate the characters of the spam messages.

The obfuscation algorithm implementation involves the following steps: (i) randomly select one word to obfuscate in each group of seven words, (ii) randomly select a character from the word to be obfuscated (iii) applying one of the possible character replacements (see Table III) and (iv) repeat the process starting from the last previously selected word until the full content of the message is processed.

TABLE II. PUBLICLY AVAILABLE SPAM DATASETS

Dataset	Content description	Spam ratio	URL
British English SMS corpora	875 SMS	48% spam	https://mtaufiqnzz.files.wordpress.com/2010/06/british-english-sms-corpora.doc
Bruce Guenter spam collection	>3,000,000 emails	100% spam	http://untroubled.org/spam/
Clueweb 09	1,040M websites (HTML)	unknown	http://www.lemurproject.org/clueweb09.php/
Clueweb 12	870M websites (HTML)	unknown	http://www.lemurproject.org/clueweb12.php/
Common Crawl Data	9 Billion in 2014 and increasing websites (HTML)	100% spam	http://commoncrawl.org/
CSDMC 2010 Spam Corpus	4327 emails	32% spam	http://csmining.org/index.php/spam-email-datasets-.html
DC 2010 / EU 2010	23M websites (HTML)	unknown	https://dms.sztaki.hu/en/letoltes/ecmlpkdd-2010-discovery-challenge-data-set
Enron email	619,446 emails	0% spam	http://www.cs.cmu.edu/~enron/
HSpam14.s2	14M Twitter messages (tweets)	unknown	https://doi.org/10.1145/2766462.2767701
Ling spam	2,893 emails	16% spam	http://csmining.org/index.php/ling-spam-datasets.html
SpamAssassin	6,047 emails	31% spam	http://spamassassin.apache.org/old/publiccorpus/
Spam Corpus	4,027 emails	34% spam	https://github.com/hexgnu/spam_filter/tree/master/data
SMS Spam Collection v.1	5,574 SMS	13% spam	https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection
TREC 2007 Public Corpus	75,419 emails	66% spam	http://plg.uwaterloo.ca/~gvcormac/treccorpus07/
Webspam-uk 2007	105,896,555 websites (HTML)	unknown	http://chato.cl/webspam/datasets/index.php
Webspam-uk 2011	3,766 Web websites (HTML)	53% spam	https://sites.google.com/site/heiderawahsheh/home/web-spam-2011-datasets/uk-2011-web-spam-dataset
Webb spam 2011	330.000 websites (HTML)	unknown	http://www.cc.gatech.edu/projects/doi/WebbSpamCorpus.html
YouTube Comments Dataset	6M Youtube comments	7% spam	http://mlg.ucd.ie/yt/
YouTube Spam Collection Dataset	1,956 Youtube comments	49% spam	https://archive.ics.uci.edu/ml/datasets/YouTube+Spam+Collection

TABLE III. EXAMPLES OF CHARACTER SUBSTITUTIONS IN *LEETSPEAK*

a	4, ^, /-, -\	n], /V, [], , (), /VV/
b	8, 3, 8, 3, 8, 13	o	0, (, [], ø
c	(, €, <, [, @, ç, €, {	p	!%, ?
d	[],),], [>,]	q	"(_)", "()",
e	€, 3, [-, .€	r	/2, 2, 12
f	=, /#, #, f	s	5, \$, \$, _/^-
g	6, (_+, , (_-	t	+, , ^-, †
h	#, /-, [-],]-[,)-(, (-), :-:, ^-, ^-,]-[-, {}	u	_], _/, (, , /_/,]_[-
i	1, !, , , , :	v	V, /
j	¿, _/, _), 7	w	VV, \^/, _/_/, _:/, ^/, '//
k	{, <, {, <	x)%, %>
l	_], []_, [_, 1_	y	'/, ¥
m	V], /^/, , (V), , [V], /VV/^\, /^^^\,	z	7_, 2, >_

When *Leetspeak* is used consciously, it is very likely that all changes applied to a particular character (e.g., "A") are always the same (e.g., "4"). However, when *Leetspeak* is used to avoid spam filters, some randomly selected characters are automatically replaced by one of its Leetspeak translations. The presented obfuscation method performs the replacements completely random using all possible replacements (see Table III).

The four datasets generated (CSDMC 2010 *Leetspeak*, TREC 2007 Public Corpus *Leetspeak*, YouTube Comments Dataset *Leetspeak*, YouTube Spam Collection Dataset *Leetspeak*) have been shared in a public repository on the website of Mondragon Unversitatea (<https://mondragon.edu>) and Zenodo [45].

III. METHODS

Our proposal involves the application of DL strategies for the identification of obfuscated characters. This section explains the identification of *Leetspeak* sequences in text (Subsection A), our proposal to decode them (Subsection B) and the experimental protocol designed for evaluation purposes (Subsection C).

A. *Leetspeak* Sequence Identification

The deobfuscation problem is identifying the character in the text that best matches a particular sequence of Leetspeak characters. The identification of *Leetspeak* sequences is done by detecting non-alpha characters included in words (sequence of characters that do not contain spaces). In particular, we search for the first and the last non-alpha characters in a word and select the characters included between them as a *Leetspeak* sequence.

TABLE IV. EXAMPLES OF OBFUSCATED LEETSPEAK CHARACTER SEQUENCES

Obfuscated character	Generated image	Identified character
_	_	L
†	†	T
€	€	E
		I
[]	[]	N

Once the obfuscated character has been detected and isolated, it is transformed into an image (see examples included in Table IV).

Then, the images are classified using neural networks (DL) for the identification of the obfuscated character. The identified character is used to rewrite the original word and the process starts again until the end of the message. Once a message has been fully decoded, its classification can be successfully performed by taking advantage of common text classification processes. The following subsection explains the DL scheme used to decode *Leetspeak* character sequences.

B. Character Identification Model

For the identification of characters, we take advantage of an image recognition system that does not require the use of static dictionaries. The recognition of each obfuscated character involves looking at some specific sequences of punctuation marks and numbers that are visually similar to the target character. The sequences used to encode a character can be of different length. For example, the 'V' character may consist of two consecutive punctuation marks (i.e. a backslash and a slash, '\V'). However, in the case of the character 'H', it is more common to use three punctuation marks (i.e. ']-['). Furthermore, our proposal should also recognize new *Leetspeak* variants used to obfuscate words. Keeping in mind these considerations, our proposal includes a CNN. Table V provides detailed information of the layers that make up the CNN design.

TABLE V. LAYER DETAILS OF OUR CNN USED FOR CHARACTER IDENTIFICATION

#	Convolution
1	Conv2D(filters 32, kernel_size (3,3), activation_function=relu, stride=(1,1) MaxPooling2D poolsize=(2,2)
2	Conv2D(filters 64, kernel_size (3,3), activation_function=relu, stride=(1,1) MaxPooling2D poolsize=(2,2)
3	Conv2D(filters 128, kernel_size(3,3), activation_function=relu, stride=(1,1) MaxPooling2D poolsize=(2,2)
	Dropout (0,7)
	Flatten ()
	Dense (neurons 512, activation_function=relu)
	Dense (neurons 26, activation_function=softmax)

As shown in Table V, we have defined our CNN as a stack of alternate layers of Convolution, ReLU and MaxPooling. The shape of the input data is a 100x100 pixels image with a colour depth of 1 byte and the output layer comprises 26 neurons and a "softmax" activation function (computes the probability of identifying a specific text character).

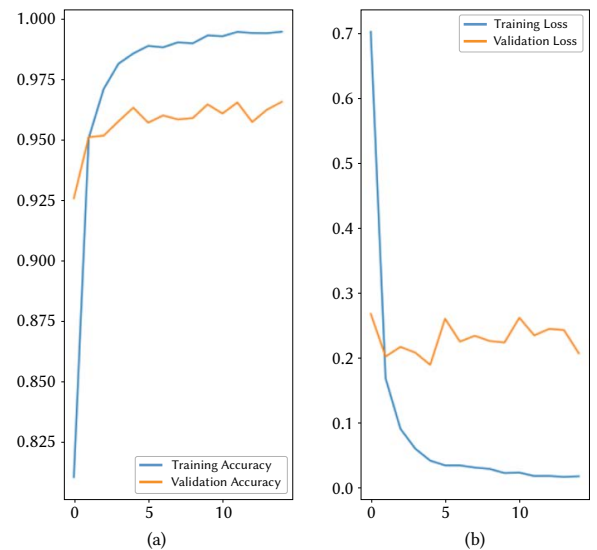


Fig. 3. Evaluation of performance achieved during Training and Validation stage. Parts: a) training and validation accuracy, b) training and validation loss.

The model has been trained over 15 epochs and 20% of the training dataset (image database described in Subsection II.B) has been reserved to validate model's performance over the different epochs. In addition, the possibility of adding an early stop as a callback in the training process has been considered to reduce overfitting. However, it was

decided to train the model on a certain number of epochs, as the model will try to predict obfuscations formed by characters, but it will only be trained with different variations of real letters. Therefore, in this case it is not essential to apply an early stop to avoid overfitting the model. The *accuracy* and *loss* measurements for training and validation are shown in Fig. 3.

Fig. 3a shows the *accuracy* obtained by the model in each epoch for training and validation datasets. Fig. 3b shows the loss evolution for each epoch. As can be seen, after a few epochs (10) we obtain an *accuracy* close to 90%. After that, the increase in *accuracy* is slower (the neural network needs many epochs to achieve small improvements in *accuracy*).

C. Experimental Protocol

To evaluate the performance of our CNN in a real environment, we created specific test datasets containing spam messages with obfuscated characters (see Subsection II.A). The evaluation was carried out using an experimental protocol (Fig. 4) designed for this purpose.

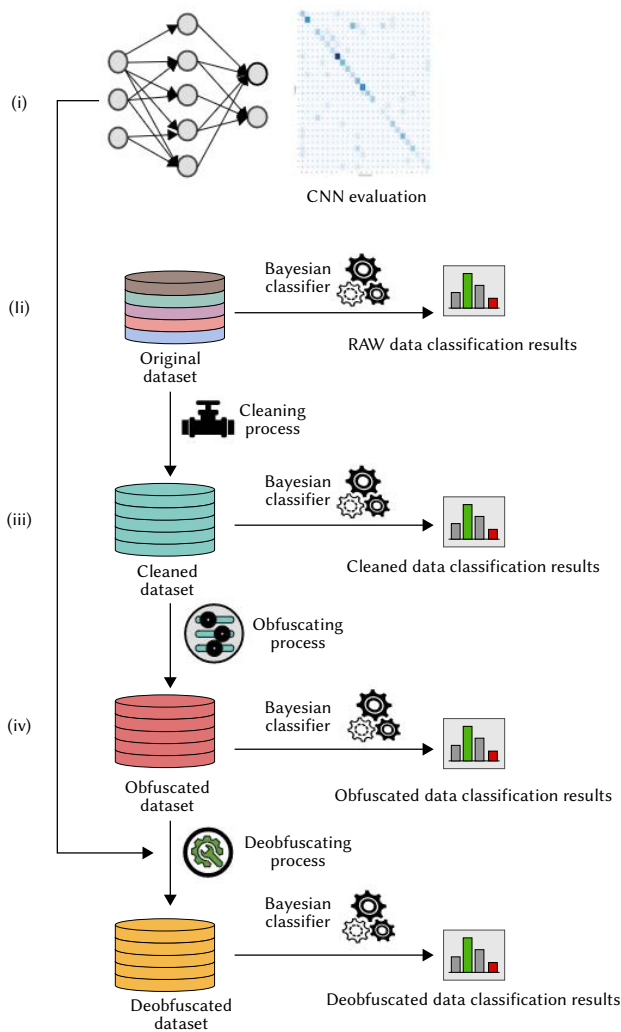


Fig. 4. Experimental protocol.

As shown in Fig. 4, the experiment comprises five steps in which different aspects are evaluated: (i) the CNN, (ii) the classification of the original datasets (baseline), (iii) the classification of the dataset after applying a cleaning process over the original datasets, (iv) the classification of the obfuscated datasets and (v) the classification of the deobfuscated datasets.

The first step involves training the CNN and evaluating its performance for Leetspeak deobfuscation. For this purpose, we used and analysed a confusion matrix generated by classifying all Leetspeak sequences included in Table III.

During the second step, messages were classified in their original form to obtain a set of baseline performance measures. Due to the large number of different classifiers, we selected the 10 best classifiers identified in the previous work of Ezpeleta *et al.* [44].

The third step consists of identifying and removing non-alphanumeric characters (text cleaning) from the dataset represented in its original form and classifying again the resulting texts. Additionally, in each message, the phone numbers and web URLs included in the message were retained and the rest of the text was converted to lowercase. Pre-processing the messages as described above, we achieved new classification results.

Finally, we classified the obfuscated datasets using the process defined in Subsection II.B (step 4) and the same datasets after being deobfuscated (step 5).

The analysis of results included a comparison of the performance achieved during the last four steps (baseline - step 2, cleaned - step 3, obfuscated - step 4 and deobfuscated - step 5) using standard measures including: *accuracy*, *precision*, *recall* and *f-score* [REF]. We have used a 10-fold cross-validation scheme to run experiments in the last four steps.

IV. RESULTS AND DISCUSSION

This section contains the results obtained in the experimentation. First, the implemented CNN was directly evaluated using a confusion matrix. The confusion matrix (Fig. 5) was generated by classifying a dataset of 115 *Leetspeak* sequences.

a	3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
b	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
c	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	1	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0		
d	0	0	0	3	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
e	0	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
f	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
g	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
h	1	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
i	0	0	0	1	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		
j	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
k	0	1	0	0	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
l	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
m	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
n	0	0	0	0	0	0	0	0	0	0	0	1	3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
o	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
p	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
s	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
t	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
u	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
v	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
w	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
x	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
z	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z											

Fig. 5. Confusion matrix achieved using our CNN.

As shown in Fig. 5, although there are some errors, the main diagonal of the confusion matrix shows a large number of hits in recognizing *Leetspeak* sequences. Our CNN has achieved the following performance scores in *Leetspeak* identification: Accuracy = 0.62, Recall = 0.62, Precision = 0.62, f1-score = 0.62, Kappa = 0.61 and Matheus_Coefficient = 0.61.

Then, a practical approach was taken to see how the accuracy of the CNN identification was carried over to the text classification phase. To this end, in fifth step (deobfuscated) the identified *Leetspeak* sequences are replaced by their translations (obtained from the CNN) and then the text is preprocessed and classified. The configurations of classifiers and preprocessing used for the classification process are described in Table VI.

TABLE VI. LEGEND FOR EXPERIMENTAL CONFIGURATIONS

Symbol	Meaning
MBM	Multinomial Naïve Bayes
MBMU	Multinomial Naïve Bayes Updateable
CNB	Complement Naïve Bayes
.c	Output Word Counts (outwc)
.c	Use a binary representation for tokens (0 1)
.stvw	String to Word Vector
.go	Using the following Weka options (-L -O -W 10000000)
.go	Using default Weka options
.ngtok	NGram Tokenizer 1-3
.ngtok	NGram Tokenizer is not used
.stemmer	Stemmer
.stemmer	Stemmer is not used

Fig. 6 shows the *accuracy* evaluations achieved using the 10 best preprocessing/classification configurations performed in our previous work [44]. The figure has been divided in four separate parts grouping all configurations done by each dataset.

As shown in Fig. 6, the best configurations are those with the original dataset (Baseline and Cleaned). However, when spammers take advantage of *Leetspeak*, using the deobfuscation scheme introduced in this work contributes to improved classification results for all datasets and preprocessing/classification configurations analysed. The use of the deobfuscation schemes allows, in some configurations, to achieve classification results close to those obtained when spammers do not obfuscate the emails (Baseline and Clean). Therefore, the use of CNNs

allows good deobfuscation results to be obtained without no need for other complex procedures.

In addition, we also performed an evaluation of the impact of our deobfuscation scheme using precision and recall measures. Table VII shows precision and recall evaluations achieved for all datasets.

As shown in Table VII, the results present the same behaviour as for the *accuracy* evaluation and confirm the utility of the deobfuscation process. Finally, we executed a *f-score* evaluation using all datasets to check whether the deobfuscation was worth according to other criteria. Results are shown in Fig. 7.

As shown in Fig. 7, the *f-score* evaluations through the different scenarios are very similar to previous evaluations obtained for *accuracy*, *recall* and *precision*. The results achieved indicate that substantial performance benefits can be obtained by a deobfuscation process based on the use of CNNs such as the one shown in this study. These successful results are due to the ease with which CNNs automatically detect important features without the need for human supervision. In addition, the use of a wide variety of fonts and styles during CNN training allowed for greater accuracy in the identifying *Leetspeak* sequences.

However, it is very important to select a suitable dataset (such as the one provided as a result of the present research) that allows CNN to learn how to decode *Leetspeak*. Next section shows the main conclusions and outlines future work.

V. CONCLUSIONS AND FUTURE WORK

This study aims to discover mechanisms for automatically decode *Leetspeak* character sequences using only CNN-based models. We provide (i) a reliable CNN design for *Leetspeak* deobfuscation processes and its evaluation, (ii) an image database that has been used for training the CNN model in this study and (iii) four datasets for evaluating *Leetspeak* deobfuscation processes. Through experimental testing, we find that the CNN design and creation processes are able to achieve great performance.

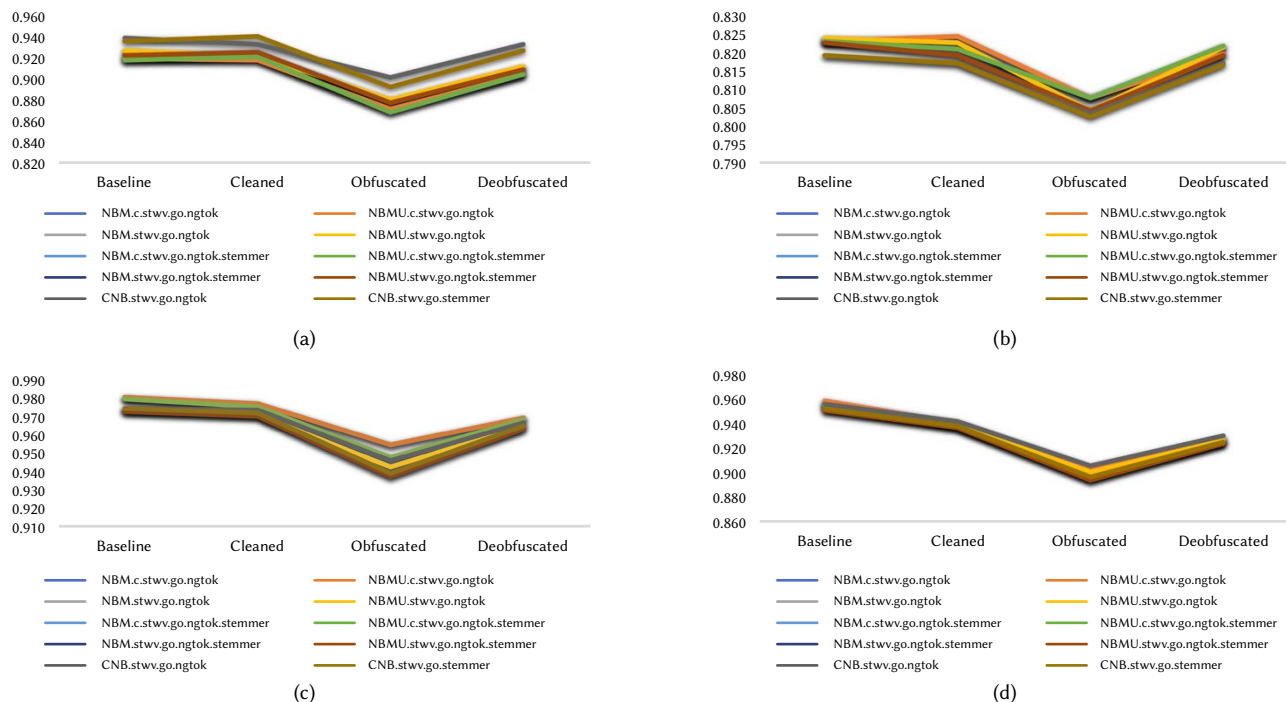


Fig. 6. Experimental results achieved using accuracy measure. Parts: a) Youtube Spam Collection, b) Youtube Comments, c) CMDMC2010 dataset, d) TREC2007 dataset.

TABLE VII. PRECISION AND RECALL EVALUATIONS FOR DATASETS

Classifier/preprocessing configuration	Dataset status	YouTube Comments Dataset		YouTube Spam Collection Dataset		CSDMC 2010		TREC 2007	
		Measure	precision	recall	precision	recall	precision	recall	precision
NBM.c.stvw.go.ngtok	Baseline	0.801	0.387	0.884	0.972	0.991	0.948	0.987	0.847
	Cleaned	0.798	0.399	0.880	0.974	0.992	0.936	0.991	0.767
	Obfuscated	0.802	0.308	0.807	0.984	0.999	0.861	0.998	0.617
	Deobfuscated	0.808	0.366	0.857	0.979	0.991	0.913	0.993	0.718
NBMU.c.stvw.go.ngtok	Baseline	0.801	0.387	0.884	0.972	0.991	0.948	0.987	0.847
	Cleaned	0.798	0.399	0.880	0.974	0.992	0.936	0.991	0.767
	Obfuscated	0.802	0.308	0.807	0.984	0.999	0.861	0.998	0.617
	Deobfuscated	0.808	0.366	0.857	0.979	0.991	0.913	0.993	0.718
NBM.stvw.go.ngtok	Baseline	0.834	0.371	0.894	0.973	0.992	0.927	0.990	0.829
	Cleaned	0.820	0.373	0.892	0.966	0.994	0.916	0.991	0.763
	Obfuscated	0.809	0.284	0.822	0.982	0.997	0.824	0.997	0.603
	Deobfuscated	0.836	0.357	0.870	0.976	0.994	0.898	0.992	0.714
NBMU.stvw.go.ngtok	Baseline	0.834	0.371	0.894	0.973	0.992	0.927	0.990	0.829
	Cleaned	0.820	0.373	0.892	0.966	0.994	0.916	0.991	0.763
	Obfuscated	0.809	0.284	0.822	0.982	0.997	0.824	0.997	0.603
	Deobfuscated	0.836	0.357	0.870	0.976	0.994	0.898	0.992	0.714
NBM.c.stvw.go.ngtok.stemmer	Baseline	0.822	0.375	0.881	0.973	0.990	0.946	0.989	0.828
	Cleaned	0.805	0.376	0.883	0.978	0.993	0.932	0.993	0.757
	Obfuscated	0.828	0.293	0.805	0.982	0.999	0.839	0.998	0.584
	Deobfuscated	0.826	0.366	0.857	0.978	0.993	0.909	0.996	0.706
NBMU.c.stvw.go.ngtok.stemmer	Baseline	0.822	0.375	0.881	0.973	0.990	0.946	0.989	0.828
	Cleaned	0.805	0.376	0.883	0.978	0.993	0.932	0.993	0.757
	Obfuscated	0.828	0.293	0.805	0.982	0.999	0.839	0.998	0.584
	Deobfuscated	0.826	0.366	0.857	0.978	0.993	0.909	0.996	0.706
NBM.stvw.go.ngtok.stemmer	Baseline	0.847	0.355	0.890	0.969	0.992	0.924	0.991	0.810
	Cleaned	0.820	0.355	0.894	0.971	0.994	0.914	0.991	0.754
	Obfuscated	0.847	0.265	0.817	0.981	0.998	0.806	0.997	0.579
	Deobfuscated	0.856	0.333	0.863	0.978	0.994	0.891	0.994	0.700
NBMU.stvw.go.ngtok.stemmer	Baseline	0.847	0.355	0.890	0.969	0.992	0.924	0.991	0.810
	Cleaned	0.820	0.355	0.894	0.971	0.994	0.914	0.991	0.754
	Obfuscated	0.847	0.265	0.817	0.981	0.998	0.806	0.997	0.579
	Deobfuscated	0.856	0.333	0.863	0.978	0.994	0.891	0.994	0.700
CNB.stvw.go.ngtok	Baseline	0.750	0.415	0.917	0.972	0.991	0.930	0.990	0.833
	Cleaned	0.742	0.412	0.915	0.959	0.994	0.923	0.991	0.777
	Obfuscated	0.734	0.337	0.853	0.976	0.997	0.835	0.997	0.625
	Deobfuscated	0.755	0.398	0.907	0.969	0.993	0.903	0.992	0.728
CNB.stvw.go.ngtok.stemmer	Baseline	0.779	0.388	0.912	0.969	0.992	0.927	0.992	0.818
	Cleaned	0.757	0.396	0.924	0.965	0.995	0.919	0.991	0.758
	Obfuscated	0.757	0.311	0.841	0.975	0.998	0.812	0.997	0.587
	Deobfuscated	0.768	0.384	0.896	0.971	0.994	0.898	0.993	0.707

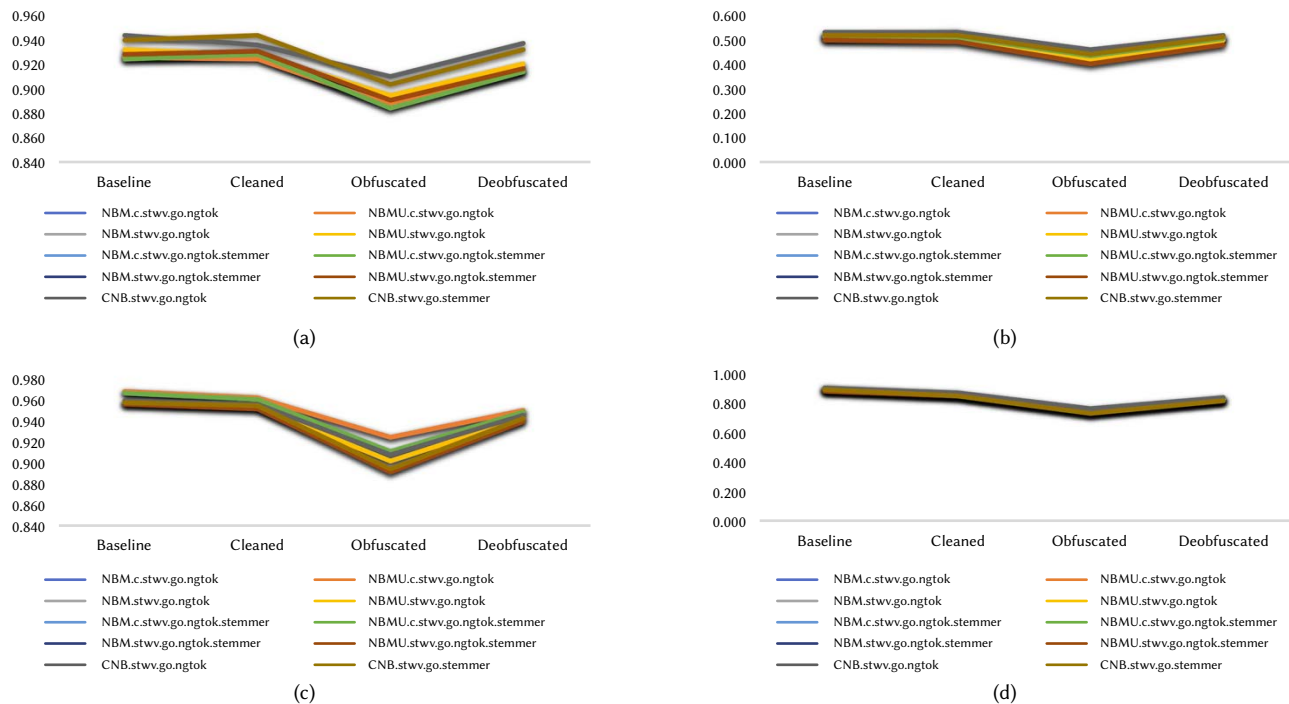


Fig. 7. Experimental results achieved using f-score measure. Parts: a) Youtube Spam Collection, b) Youtube Comments, c) CMDMC2010 dataset, d) TREC2007 dataset.

Analysing the classification rates from the clean text, we can conclude that using *Leetspeak* schemes to obfuscate characters has a huge impact on the performance of all algorithms. By obfuscating characters, spammers are able to completely hide words and make them unusable in spam classification processes. When messages are deobfuscated, the performance of the classifiers increases and reaches, in many cases, the values obtained when messages have not been obfuscated. This fact demonstrates that our proposal can be successfully used to identify the obfuscated characters. However, as shown in Fig. 5, some characters are not correctly identified and further improvements are necessary. Therefore, future work includes extending the image database and improving the CNN architecture to obtain better deobfuscation results.

The main limitation of our proposal is the detection of obfuscated characters containing one single punctuation mark because this requires further analysis. For example, the character H could be obfuscated with a middle hyphen (“-”) between two “i” (i.e. “i-i”). This situation could lead to a large number of decoding errors (e.g. “semi-interlaced” being translated into “semhnterlaced”, which is incorrect). To address this problem, we consider using dictionary-based schemes (to search whether the word exists with no changes) before using a deobfuscation algorithm. Additionally, we take advantage of the multiple outputs of the CNN (e.g. we consider the five CNN outputs that achieve the highest value) and check the existence of the resulting word in a dictionary. Moreover, the algorithm used to recognize *Leetspeak* sequences also needs to be improved. The one used in this study can only detect one *Leetspeak* sequence per word. Therefore, future work involves improving in several directions (CNN performance, algorithms to detect *Leetspeak* sequences and use of a dictionary) that will lead to significant improvements in the deobfuscation process.

ACKNOWLEDGMENT

Iñaki Velez de Mendizabal, Enaitz Ezpeleta and Urko Zurutuza are part of the Intelligent Systems for Industrial Systems research group of

Mondragon Unibertsitatea (IT1676-22), supported by the department of Education, Universities and Research of the Basque Country.

We are supported by the project Semantic Knowledge Integration for Content-Based Spam Filtering, subprojects TIN2017-84658-C2-1-R and TIN2017-84658-C2-2-R, from SMEIC, SRA and ERDF.

Vitor Basto Fernandes acknowledges FCT – Fundação para a Ciência e a Tecnologia, I.P., for its support in the context of project UIDB/04466/2020 and UIDP/04466/2020.

REFERENCES

- [1] M. Chakraborty, S. Pal, R. Pramanik, and C. Ravindranath Chowdary, “Recent developments in social spam detection and combating techniques: A survey,” *Information Processing and Management*, vol. 52, no. 6, pp. 1053–1073, 2016, doi: 10.1016/j.ipm.2016.04.009.
- [2] S. Suryawanshi, A. Goswami, and P. Patil, “Email Spam Detection: An Empirical Comparative Study of Different ML and Ensemble Classifiers,” in *Proceedings of the 2019 IEEE 9th International Conference on Advanced Computing '19*, Tiruchirappalli, India, 2019, pp. 69–74. doi: 10.1109/IACC48062.2019.8971582.
- [3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *Proceedings of the 1st International Conference on Learning Representations '13*, 2013. [Online]. Available: <http://arxiv.org/abs/1301.3781>
- [4] Y. Cabrera-León, P. García Báez, and C. P. Suárez-Araujo, “Non-email spam and machine learning-based anti-spam filters: Trends and some remarks,” in *Proceedings of the Conference on Computer Aided Systems Theory '17*, 2018, vol. 10671, pp. 245–253. doi: 10.1007/978-3-319-74718-7_30.
- [5] Z. Liu, W. Lin, N. Li, and D. Lee, “Detecting and filtering instant messaging spam - a global and personalized approach,” in *Proceedings of the 1st IEEE ICNP Workshop on Secure Network Protocols '05*, 2005, pp. 19–24. doi: 10.1109/NPSEC.2005.1532048.
- [6] C. Manning, P. Raghavan, and H. Schütze, “Introduction to information retrieval,” *Natural Language Engineering*, vol. 16, no. 1, pp. 100–103, 2010.
- [7] E. Alpaydin, *Introduction to machine learning*. Cambridge, Massachusetts: MIT press, 2020.
- [8] J. Hovold, “Naive Bayes Spam Filtering Using Word-Position-Based Attributes,” presented at the Second Conference on Email and Anti-Spam

- CEAS-2005, California, USA, 2005. [Online]. Available: <http://www.ceas.cc/papers-2005/144.pdf>
- [9] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with naive bayes-which naive bayes?," in *Proceedings of the 3rd Conference on Email and Anti-Spam*, 2006, pp. 28–69. [Online]. Available: <http://www.ceas.cc/2006/listabs.html#15.pdf>
- [10] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C. D. Spyropoulos, and P. Stamatopoulos, "Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach." arXiv cs.CL/0009009, 2000. [Online]. Available: <https://arxiv.org/pdf/cs/0009009.pdf>
- [11] S. Goyal, R. Chauhan, and S. Parveen, "Spam detection using KNN and decision tree mechanism in social network," in *Proceedings of the 4th International Conference on Parallel, Distributed and Grid Computing '16*, Himachal Pradesh, India, 2016, pp. 522–526.
- [12] S. K. Trivedi and P. K. Panigrahi, "Spam classification: a comparative analysis of different boosted decision tree approaches," *Journal of Systems and Information Technology*, vol. 20, no. 3, pp. 298–320, 2018, doi: 10.1108/JSIT-11-2017-0105.
- [13] Q. Wang, Y. Guan, and X. Wang, "SVM-Based Spam Filter with Active and Online Learning," in *Proceedings of the 15th Text REtrieval Conference*, Gaithersburg, Maryland, 2006, p. 36. [Online]. Available: <https://trec.nist.gov/pubs/trec15/papers/hit.spam.final.pdf>
- [14] J. Clark, I. Koprinska, and J. Poon, "A neural network based approach to automated e-mail classification," in *Proceedings International Conference on Web Intelligence '03*, Halifax, NS, Canada, 2003, pp. 702–705. doi: 10.1109/WI.2003.1241300.
- [15] J. Goodman and W. Yih, "Online Discriminative Spam Filter Training," in *Proceedings of the 3rd Conference on Email and Anti-Spam*, Mountain View, California, 2006, pp. 1–4. [Online]. Available: <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/GoodmanYih-ceas06.pdf>
- [16] T. Oda and T. White, "Increasing the accuracy of a spam-detecting artificial immune system," in *Proceedings of the 2003 Congress on Evolutionary Computation '03*, Camberra, Australia, 2003, vol. 1, pp. 390–396.
- [17] X. Carreras and L. Marquez, "Boosting trees for anti-spam email filtering." arXiv cs/0109015, 2001. [Online]. Available: <https://arxiv.org/abs/cs/0109015>
- [18] C. Fellbaum, "WordNet," in *The Encyclopedia of Applied Linguistics*, C. Chapelle, Ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2012, pp. 1–8. doi: 10.1002/9781405198431.wbeal1285.
- [19] R. Navigli and S. P. Ponzetto, "BabelNet: Building a very large multilingual semantic network," in *Proceedings of the 48th annual meeting of the association for computational linguistics*, Uppsala, Sweden, 2010, pp. 216–225.
- [20] J. R. Méndez, T. R. Cotos-Yañez, and D. Ruano-Ordás, "A new semantic-based feature selection method for spam filtering," *Applied Soft Computing*, vol. 76, pp. 89–104, 2019, doi: 10.1016/j.asoc.2018.12.008.
- [21] E. M. Bahgat and I. F. Moawad, "Semantic-Based Feature Reduction Approach for E-mail Classification," in *Proceedings of the 2nd International Conference on Advanced Intelligent Systems and Informatics '16*, Cairo, Egypt, 2017, pp. 53–63. doi: 10.1007/978-3-319-48308-5_6.
- [22] I. Vélez de Mendizabal, V. Basto-Fernandes, E. Ezpeleta, J. R. Méndez, and U. Zurutuza, "SDRS: A new lossless dimensionality reduction for text corpora," *Information Processing & Management*, vol. 57, no. 4, p. 102249, 2020, doi: 10.1016/j.ipm.2020.102249.
- [23] M. Dredze, R. Gevaryahu, and A. Elias-Bachrach, "Learning Fast Classifiers for Image Spam," in *Proceedings of the 3rd Conference on Email and Anti-Spam '07*, Mountain View, California, 2007, pp. 1–9. [Online]. Available: https://www.cs.jhu.edu/~mdredze/publications/image_spam_ceas07.pdf
- [24] A. Chaudhuri, K. Mandaviya, P. Badelia, and S. K. Ghosh, "Optical character recognition systems," in *Optical Character Recognition Systems for Different Languages with Soft Computing*, Cham, Switzerland: Springer, 2017, pp. 9–41.
- [25] B. Biggio, G. Fumera, I. Pillai, F. Roli, and R. Satta, "Evading SpamAssassin with obfuscated text images," 2007. <https://www.virusbulletin.com/virusbulletin/2007/11/evading-spamassassin-obfuscated-text-images/> (accessed Jun. 07, 2023).
- [26] J. Evershed and K. Fitch, "Correcting noisy OCR: Context beats confusion," in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 2014, pp. 45–51.
- [27] E. Bursztein, M. Martin, and J. Mitchell, "Text-based CAPTCHA strengths and weaknesses," in *Proceedings of the 18th ACM conference on Computer and communications security '11*, Chicago, Illinois, USA, 2011, pp. 125–138. doi: 10.1145/2046707.2046724.
- [28] J. Wang, J. Qin, X. Xiang, Y. Tan, N. Pan, and College of Computer Science and Information Technology, Central South University of Forestry and Technology, 498 shaoshan S Rd, Changsha, 410004, China, "CAPTCHA recognition based on deep convolutional neural network," *Mathematical Biosciences and Engineering*, vol. 16, no. 5, pp. 5851–5861, 2019, doi: 10.3934/mbe.2019292.
- [29] F.-L. Du, J.-X. Li, Z. Yang, P. Chen, B. Wang, and J. Zhang, "CAPTCHA Recognition Based on Faster R-CNN," in *Proceedings of the 13th International Conference on Intelligent Computing Theories and Application '17*, Liverpool, UK, 2017, vol. 10362, pp. 597–605. doi: 10.1007/978-3-319-63312-1_52.
- [30] E. Flamand, "Deciphering L33t5p34k Internet Slang on Message Boards," Diss. Ghent University, 2008. [Online]. Available: <https://lib.ugent.be/en/catalog/rug01:001414289>
- [31] J. A. Zdziarski, *Ending spam: Bayesian content filtering and the art of statistical language classification*. San Francisco, California: No starch press, 2005.
- [32] A. Tundis, G. Mukherjee, and M. Mühlhäuser, "Mixed-code text analysis for the detection of online hidden propaganda," in *Proceedings of the 15th International Conference on Availability, Reliability and Security '20*, Dublin, Ireland, 2020, pp. 1–7. doi: 10.1145/3407023.3409211.
- [33] F. K. Dosiilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," in *Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics*, Opatija, Croatia, 2018, pp. 210–215. doi: 10.23919/MIPRO.2018.8400040.
- [34] A. Tundis, G. Mukherjee, and M. Mühlhäuser, "An Algorithm for the Detection of Hidden Propaganda in Mixed-Code Text over the Internet," *Applied Sciences*, vol. 11, no. 5, Article ID: 2196, 2021, doi: 10.3390/app11052196.
- [35] T. E. de Campos, B. R. Babu, and M. Varma, "Character recognition in natural images," in *Proceedings of the 4th International Conference on Computer Vision Theory and Applications '09*, Lisbon, Portugal, 2009, pp. 273–280.
- [36] M. Deore and 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, doi: 10.9781/ijimai.2021.09.005.
- [37] A. Bhaik, V. Singh, E. Gandotra, and D. Gupta, "Detection of Improperly Worn Face Masks using Deep Learning – A Preventive Measure Against the Spread of COVID-19," *International Journal of Interactive Multimedia and Artificial Intelligence*, pp. 14–25, 2021, doi: 10.9781/ijimai.2021.09.003.
- [38] A. Jan and G. M. Khan, "Real World Anomalous Scene Detection and Classification using Multilayer Deep Neural Networks," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 2, pp. 158–167, 2021, doi: 10.9781/ijimai.2021.10.010.
- [39] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, ArticleID 7068349, 2018, doi: 10.1155/2018/7068349.
- [40] A. Ioannidou, E. Chatzilari, S. Nikolopoulos, and I. Kompatsiaris, "Deep Learning Advances in Computer Vision with 3D Data: A Survey," *ACM Computing Surveys*, vol. 50, no. 2, pp. 1–38, 2018, doi: 10.1145/3042064.
- [41] M. Abadi et al., "Tensorflow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation '16*, Savannah, GA, USA, 2016, pp. 265–283. Accessed: Mar. 24, 2022. [Online]. Available: <https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi>
- [42] A. Gulli and S. Pal, *Deep learning with Keras*. Birmingham, UK: Packt Publishing Ltd, 2017.
- [43] I. Vélez de Mendizabal, X. Vidriales, V. B. Fernandes, E. Ezpeleta, J. R. Méndez, and U. Zurutuza, "Image dataset to train a deep learning model to decode Leetspeak obfuscated characters." Zenodo, Mar. 21, 2022. doi: 10.5281/ZENODO.6373558.
- [44] E. Ezpeleta, M. Iturbe, I. Garitano, I. V. de Mendizabal, and U. Zurutuza, "A mood analysis on youtube comments and a method for improved social spam detection," in *Proceedings of the 13th International Conference*

on *Hybrid Artificial Intelligence Systems '18*, Oviedo, Spain, 2018, pp. 514–525.

- [45] I. Vélez de Mendizabal, X. Vidriales, V. B. Fernandes, E. Ezpeleta, J. R. Méndez, and U. Zurutuza, “Set of obfuscated spam dataset by using LeetSpeak transformations.” Zenodo, Mar. 21, 2022. doi: 10.5281/ZENODO.6373653.



Iñaki Vélez de Mendizabal

He is a lecturer and researcher in Mondragon Unibertsitatea. He obtained his degree in Computer Engineering from the University of Mondragon in 2003 and now he is working in his applied engineering PhD. His main research interests are in the areas of computer security and computer networks. At the present his activity focuses on designing developing and implementing semantic analysis systems for improving

antispam filters, based on the use of the Multiobjective Evolutionary Algorithms and Deep Learning.



Xabier Vidriales

He was born in the Basque Country (Spain) in 2000. Currently, he is studying Master Degree in Data Analysis, Cybersecurity and Cloud Computing at Mondragon Unibertsitatea and he has been working for three years as a research assistant in the Artificial Intelligence department at the university. He has participated in different projects and worked in several research teams, developing and

applying technologies related to spam-filtering, natural language processing and data analysis. His main interests are research on new technologies, especially those related to artificial intelligence, and the development and improvement of data analysis systems.



Vitor Basto-Fernandes

He graduated in information systems in 1995 and got his PhD on multimedia transport protocols in 2006 from University of Minho (Portugal). From 2005 he has been university lecturer, being currently assistant professor with habilitation at ISCTE – University Institute of Lisbon, Portugal. He coordinated a MSc Program in Mobile Computing and was head of the Research Centre in

Computer Science and Communications at Polytechnic Institute of Leiria. He participated in several national and international projects in information systems integration, anti-spam filtering and multi-objective optimization, publishes regularly in top-tier journals and conferences, and organized international events in the areas of his research interests, multi-objective optimization, information security and semantic web.



Enaitz Ezpeleta

Dr. Enaitz Ezpeleta is a researcher in the Data Analysis and Cybersecurity Research Group at Mondragon University. He obtained his PhD regarding New approaches for content-based analysis towards Online Social Network spam detection from Mondragon Unibertsitatea in 2016. He is the responsible for coordinating the Artificial Intelligence knowledge area at Mondragon Unibertsitatea and the

Master Degree in Data Analysis, Cybersecurity and Cloud Computing. His main research interest applies to spam filtering, Natural Language Processing and data analysis for security. During the last years, Enaitz published over a dozen journal and conference articles and has participated and coordinated work packages and tasks in different public funded research projects, including European, Spanish and Basque Government funded ones.



José R. Méndez

He was born in Galicia (Spain) in 1977. Currently, he works at the computer science department of University of Vigo as associate professor. He worked as a system administrator, software developer, and IT (Information Technology) consultant in civil services and industry during 10 years. He is an active researcher belonging to SING group and, although collaborates in different applications machine

learning, his main interests are the development and improvement of anti-spam filters. (<http://sing-group.org/>).



Urko Zurutuza

Dr. Urko Zurutuza is the principal investigator of the Intelligent Systems for Industrial Systems research group, recognised by the Basque Government Department of Education as a Type A (highest qualification). He coordinates the research activities of this group. He obtained his PhD in January 2008 at Mondragon Unibertsitatea, in a thesis carried out in collaboration with Zürich IBM research

Lab (with a pre-doctoral grant from the Basque Government). He has been lecturing in different Telecommunications and Computer Engineering Degrees and Masters, and at the PhD Programme in Applied Engineering at Mondragon Unibertsitatea. He has supervised 8 doctoral theses, and has 3 in progress. He has published more than 20 articles in high impact journals, more than 55 publications in blind peer-reviewed conferences, edited 3 books (2 of them as conference proceedings), and coauthored 7 book chapters. He has experience in 7 European projects funded under programmes such as H2020, CEF Telecom, ECSEL or ARTEMIS. He is also active in knowledge transfer activities, leading more than 40 projects with companies. He has been responsible for International Relations for Telecommunications Engineering, Computer Engineering, and Embedded Systems Master degrees between 2010 and 2018. He is member of the Academic Committee of the Doctoral Programme since 2012. He is member of the Board of Directors of the National Network of Excellence in Cybersecurity Research, and Steering Board member of leading international scientific conferences such as DIMVA or RAID.

Attentive Flexible Translation Embedding in Top-N Sparse Sequential Recommendations

Min-Ji Seo, Myung-Ho Kim *

Department of Software Convergence, Soongsil University, 369, Sangdo-ro, Dongjak-gu, Seoul 06978 (Korea)

Received 9 September 2021 | Accepted 2 September 2022 | Published 19 October 2022



ABSTRACT

Sequential recommendation aims to predict the user's next action based on personal action sequences. The major challenge in this task is how to achieve high performance recommendation under the data sparsity problem. Translation-based recommendations, which learn distance metrics to capture interactions between users and items in sequential recommendations, are a promising method to overcome this issue. However, a disadvantage of translation-based recommendations is that they capture long-term preferences of the user and complex item transitions. In this paper, we propose attentive flexible translation for recommendations (AFTRec) to tackle data sparsity problem by capturing a user's dynamic preferences and complex interactions between items in user's purchasing behaviors. In particular, we first encode semantic information of an item related to user's purchasing behaviors as the user-specific item translation vectors. We also design a transition graph and encode complex item transitions as correlation-specific item translation vectors. Finally, we adopt a flexible distance metric that considers directions with respect to the translation vectors in the same space for predicting the next item. To evaluate the performance of our method, we conducted experiments on four sparse datasets and one dense dataset with different domains. The experimental results demonstrate that our proposed AFTRec outperforms the state-of-the-art baselines in terms of normalized discounted cumulative gain and hit rate on sparse datasets.

KEYWORDS

Deep Learning, Gated Graph Neural Network, Knowledge Graph Embedding, Recommender Systems, Self-Attention, Sequential Recommendation.

DOI: 10.9781/ijimai.2022.10.007

I. INTRODUCTION

RECOMMENDER systems (RSs) have received interest on various platforms, such as e-commerce, news portals, and social media sites. The main purpose of RSs is to suggest the most relevant recommendations to users so that they make informed purchasing choices. Traditional recommendation systems [1]–[3], such as collaborative filtering (CF), make recommendations by analyzing historical interactions or preferences based on the similarity of users or items in the past. However, following the explosive growth of e-commerce, the data sparsity problem, which refers to the difficulty in finding sufficient similar users and items due to insufficient user-item interactions, is the main challenge in RS. To address this issue, matrix factorization (MF) [4] models that map both the user and item embedding vectors and represent user-item interactions by the inner product of the user and item vectors have been proposed.

To deal with sequential user behaviors (e.g., click and purchase) in e-commerce, sequential recommendation systems [5]–[8] have been proposed in RS for data sparsity problems. Examples of such models include factorized personalized Markov chains (FPMC) [9], which combine Markov chains (MCs) [10] and MF to predict the next action

of the user in sequential data. The FPMC captures both long-term user preferences and short-term sequential dynamics by modeling the interactions between user-to-item and item-to-item pairs. This underlies personalized MCs, where a user-specific transition matrix is applied to capture personalized item transitions. Achieving better performances on sparse datasets, many researchers have recently found new ways to capture interactions between user-to-item and item-to-item pairs.

Translation-based methods [11]–[14], which facilitate knowledge graph (KG) completion [15]-based approaches, such as translation-based recommendation (TransRec) [16], latent relational metric learning (LRML) [17], and collaborative metric learning (CML) [18], have achieved high performance with sparse datasets for next item recommendation. TransRec utilizes KG completion to model users as translation vectors from their previously purchased item vectors to the vector of the next items in the same translation space. To model item-to-item interactions in chronological order, TransRec adopts a translational principle, which minimizes the distance between the translation vectors. However, these translation-based recommendation methods have several drawbacks in sequential recommendations. First, they mainly adopt translating embeddings for modeling multi-relational data (TransE) [19], which is capable of 1-to-1 relations but is unable to handle 1-to-N, N-to-1, and N-to-N relations. Second, there are few studies that focus on the user's long-term and short-term preferences in translation vectors without user and item context information, such as category and user profile.

* Corresponding author.

E-mail addresses: porito2@soongsil.ac.kr (M. J. Seo), kmh@ssu.ac.kr (M. H. Kim).

In this study, we propose an attentive flexible translation for recommendations (AFTRec) to predict the user’s next item in sparse sequential recommendation datasets for the data sparsity problem. Specifically, unlike existing approaches, which primarily focus on the last consumed item, we focus on the sequential behaviors of the user and complex interactions between purchased items by users in chronological order. To facilitate KG completion in predicting the next item, we generate user-specific item translation vectors that reflect dynamic user preferences and target item translation vectors that represent the user’s next item as entities. We also generated a correlation-specific item translation vector that reflects item-to-item interactions in user behavior histories as a relation vector. For KG completion to predict the next item, we propose a distance function that can flexibly handle not only 1-to-1, but also 1-to-N, N-to-1, and N-to-N relations. Our AFTRec consists of three modules: a user-specific item translation vector embedding module, correlation-based item translation vector embedding module, and attentive item translation vector embedding module. The model applies KG completion to translation vectors for moving the user’s previous item vectors close to the user’s next item vectors in the same translation space. First, we generated item embeddings based on user behaviors through a self-attention mechanism, which is efficient for capturing long-term item dependencies with the position information of the item. For the correlation-specific item translation vector, we initially designed the transaction graph and linked the neighbors of items based on a sliding window, which slides the item sequences in a window-by-window manner. Then, we learned item-to-item interactions in sequential user behaviors by utilizing gated graph neural networks (GGNNs) [20], which are capable of representing sophisticated item interactions with comprehensive item transitions in user behavior sequences. In the final module, we generate the attentive item translation vector that aggregates the user’s sequential preferences and relationships between items and then embed translation vectors into the same space with KG embedding for translation from a previous item to the next item. Inspired by the flexible translation (FT) [21] of KG embedding, we designed our translational distance function in a new manner to model translation vectors. Therefore, unlike other existing translation-based RSs for sparse sequential datasets, AFTRec can capture not only personal item preferences but also sophisticated item interactions based on users’ and users’ sequential behaviors.

Our contributions can be summarized as follows:

1. We propose a novel translation-based sequential recommendation model. We adopt the KG embedding technique to encode sequential behaviors of a user and item-to-item relationships as entities and relations of a KG triple. We model various correlations between entities and relations to find the next item with our translational distance function, which releases existing translation approaches. Using this approach, AFTRec can capture pairwise relations between users and items more efficiently.
2. We embed sequential user preferences as a user-specific item translation vector as the head entity by applying a self-attention mechanism [22] in chronological order to understand long-term user preferences. For the secondary head entity of our distance function, we define the attentive item translation vector. The attentive item translation vector summarizes the item correlations related to the purchasing preference of each user through the soft-attention mechanism. Hence, we consider various perspectives on user and item information to translate the previous item into the next item.
3. We represent item-to-item interactions as correlation-specific item translation vectors as a relation of the KG triple through GGNN. Initially, we design a transaction graph by connecting adjacent items in chronological order using a sliding window

method. In particular, we divide edges into incoming and outgoing edges, and thus efficiently represent item interactions with the purchase order in terms of the time position. In addition, we utilize the GGNN to analyze item interactions. Because the GGNN uses a gated recurrent unit (GRU) [23] as an updater, it helps reduce the number of parameters for analysis.

4. We conduct extensive experiments using four sparse datasets and one dense dataset from different domains to evaluate the proposed method. The experimental results demonstrate that our method outperforms other existing approaches in solving the data sparsity problem.

The remainder of this paper is organized as follows. Related studies are introduced in Section II. Next, we describe our proposed method in Section III. In Section IV, we describe the experiments conducted on publicly available datasets of several domains, evaluate our proposed method in comparison with other approaches, and analyze the experimental results. Finally, we conclude the paper in Section V.

II. RELATED WORK

A. Traditional Recommender Systems

RSs aim to predict user preferences and suggest relevant items to the users. Traditionally, CF-based methods are used in RSs. CF recommends items in which similar users are interested based on historical data. For example, MF models the explicit feedback of a user with user and item latent factors by calculating the dot product of the two latent factors. To address item-based CF, a factored item similarity model [24] embeds each item and models the similarity between two items using the inner product of their embedding vectors. The neural attentive item similarity model [25] assigns an attentive weight to each item in the item sequences and shows good results in the calculation of the similarity between items. Bobadilla et al. [26] utilize neural CF to obtain prediction reliabilities and combine the prediction value and the reliability information in user ratings. Bobadilla et al. [27] improve fairness in RSs by combining probabilistic MF and multi-layer network. However, these methods have limitations in handling sequential patterns in interactions between users and items.

B. Sequential Recommender Systems

Sequential RSs investigate sequential behaviors of a user to recommend the next item. MC-based methods temporarily capture item transitions and predict the next item based on the last consumed item. FPMC fuses MF and MC to predict the next actions with the user’s general interests and short-term item transitions.

Inspired by the success of neural networks, various neural-network-based methods have been introduced for sequential RSs. Recurrent neural network (RNN)-based recommendations [28], [29] employ variations of RNN, such as long short-term memory and GRU, which are capable of modeling sequential patterns, to predict the next action of the user. However, because RNN-based recommendations contain information regarding the final state of the model, they are limited in modeling long sequences. To address this problem, attention-based RNN methods [30], [31] have been proposed. A neural attentive recommendation machine [32] applies an attention mechanism to a stacked GRU-based encoder–decoder to model the sequential behavior and capture general preferences of the user. Recently, self-attention mechanisms have become popular, with promising performance in natural language processing (NLP) problems. Accordingly, many researchers have utilized self-attention to provide suitable recommendations in historical sequences. Self-attentive sequential recommendation (SASRec) [33] uses stacked self-attention blocks to efficiently consider long-term dependencies. A stochastic shared

embeddings-personalized transformer (SSE-PT) [34] introduced personal information into self-attention by concatenating item embedding and embedding it into the self-attention embedding layer. Time interval-aware self-attention-based sequential recommendation (TiSASRec) [35] utilizes both the absolute positions of the items and the time intervals between items in a sequence. It represents the relationship between items as a time interval and shows performance improvement on a personalized sequential recommendation using two types of item positions: sequential and relative time positions.

However, these methods have several limitations. First, MC-based approaches predict the next item using only the last consumed item; thus, they do not explicitly capture the complex and long-term dependencies. Second, convolutional neural networks [36], [37] and RNN-based methods involve the risk of missing crucial information on previously consumed items and lack explanations for recommended items. Third, self-attention-based methods have insufficient ability to treat complex item-to-item and user-to-item interactions.

C. Translation-Based Embedding Model in Recommender Systems

The goal of KG completion-based RSs is to learn the relationships between users and items by minimizing the distance between the translation vectors in the same space. Fig. 1 describes TransE embedding and two KG-based approaches in RSs, namely, CML and LRML. KG-based recommendation methods initially utilize KG embeddings to predict user’s item ratings or implicit next interactions between users and items in RSs. CML minimizes the distance between the user and item vectors using personalized historical implicit feedback. LRML uses a memory-based attention network to represent the latent relationships between the user and previous item vectors as latent relation vectors. Then, LRML advances the metric learning of CML, which operates via $p \approx q$ to $p + r_1 \approx q$, where p and q are the user and item vectors, respectively, and r_1 is a user-to-item relation vector. CML and LRML then apply the model’s distance function to find the next item vector with the shortest distance from the user vector.

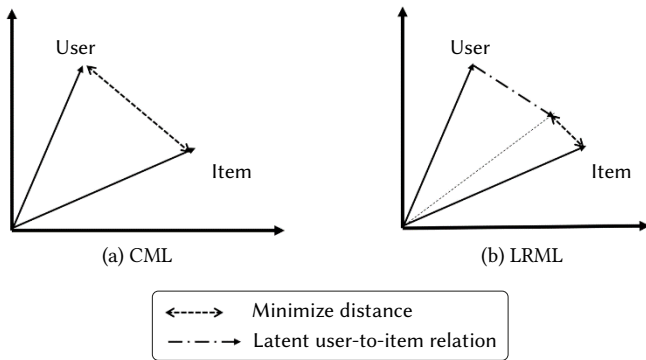


Fig. 1. Simplified illustration of (a) CML and (b) LRML.

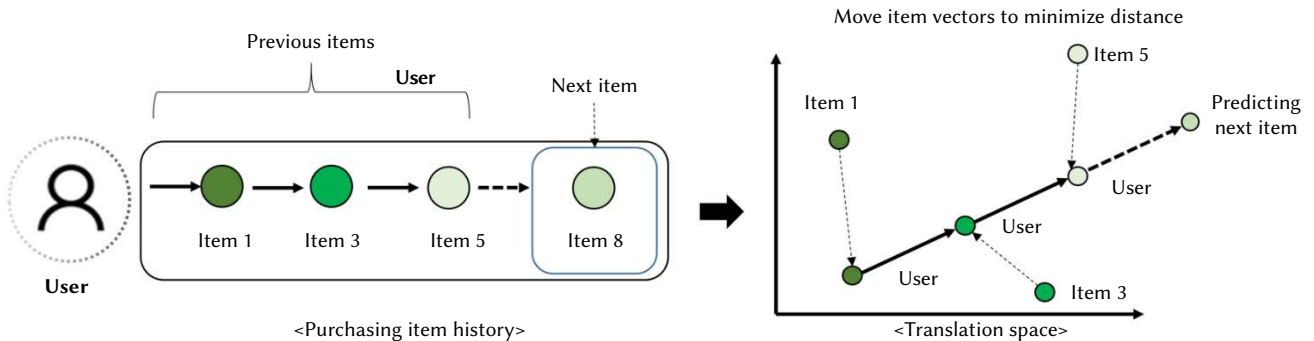


Fig. 2. Simplified illustration of TransRec based on the user’s item sequences.

To solve the data sparsity problem in sequential RSs, translation-based recommendations have been proposed. Translation-based recommendations embed the user and item vectors as translation vectors of a triple form (head, relation, tail) using KG completion to move the previous user’s item vector to be close to the user’s next item vector in sequential behavior sequences. Fig. 2 illustrates the process of providing the next item recommendation through TransRec. In Fig. 2, TransRec predicts the user’s next item by modeling third-order interactions between the user’s previous item, the user, and the user’s next item in a translation space. The previous items of the user are also modeled to move the previous item to the next item in chronological order through TransE, as shown in Fig. 2. Fig. 3 illustrates the process of providing the next item recommendation through mixtures of heterogeneous recommenders (MoHRs) [38]. Similar to TransRec in Fig. 2, MoHR predicts the user’s next item by modeling third-order interactions based on the user’s item sequences and user information. Specifically, MoHR represents various sequential relationships, that is, previous item-to-next item and user-to-next item, and adopts KG embedding to predict the next item-based distance from the previous item vector in the translation space. MoHR captures three types of relationships: long-term user-to-item preferences, relationships between short-term item transitions, and exhibit relationships (e.g., also-bought/also-viewed) between the short-term item transitions by applying TransE separately for each relationship, as shown in Fig. 3. MoHR also models item vectors in the user’s purchasing sequences to move the previous item vector close to the user’s next purchased item vector. An attentive translation model for next item recommendation (ATM) [39] constructs a user, multiple previous items, and the next items as translation vectors to translate a user to the next item. In particular, ATM implements high-order MCs to embed a user’s sequential behaviors into the relation vector. ATM then models third-order interactions (a user, the user’s sequential preferences, and the next item).

Recently, translation-based recommendations have also facilitated KG completion to predict user-to-item ratings in sequential RSs [11], [14]. Translation-based factorization machines [40] combine KG completion and factorization machines to predict user-item ratings in sequential RSs. To improve the performance of translation-based recommendations, recent approaches utilize user and item context information, such as item category and user region. The adaptive hierarchical translation-based sequential recommendation [41] captures item sequence patterns based on implicit purchasing behaviors and purchased item category information by modeling sequential item interactions using KG completion.

However, these approaches require additional resources and time to consider contextual attributes. To reduce the resources of context analysis, we solved the data sparsity problem and predicted the next item using only implicit interactions between the user and items inspired by TransRec and MoHR with performance comparable to recent sequential recommendation methods such as SSE-PT and TiSASRec introduced in Section II.B.

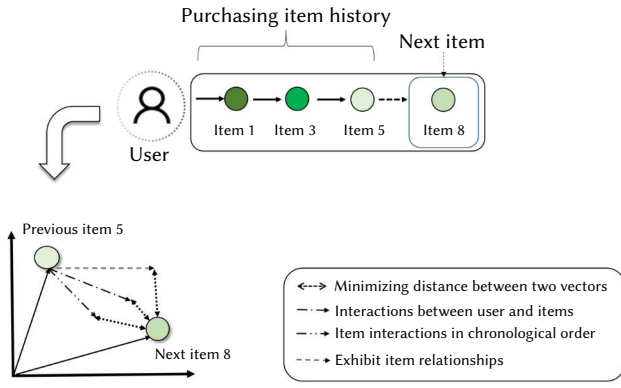


Fig. 3. Simplified illustration of MoHR. MoHR models distance functions for three types of relationships to each previous item pairs. Then, MoHR finds the item vector with the shortest distance from the last purchased item as the next item.

Many translation-based recommendation systems exhibit robust performance on sparse datasets, such as e-commerce, by adopting the translational distance for capturing third-order interactions (a user, a previously consumed item, the next item) to the next item recommendation. TransE models $h + r \approx t$, where (h, r, t) is a triple of KG, with a promising performance in 1-to-1 relationships, but it is too strict to model 1-to-n, n-to-1, and n-to-n relationships. In addition, many approaches cannot address users' long-term dependencies and thus achieve lower performance than RNN- and self-attention-based sequential recommendation systems in sequential recommendation. To address this issue, our proposed method uses a translational embedding model that handles not only 1-to-1 and other relationships but also long-term dependencies and sophisticated item interactions in sequential behaviors to recommend the most appropriate target item.

III. METHODOLOGY

In recommendation research, many translation-based approaches have been proposed that learn the relationships between users and items as translation vectors for sequential recommendation. In this

section, we introduce the novel translation-based recommendation AFTRec, which applies KG embedding to improve the sequential recommendation with sparse datasets. The architecture of the proposed AFTRec is shown in Fig. 4. First, we encoded the information of a user's consumed item based on personalized sequential behaviors to the user-specific item translation vector γ_u as the head entities. In this process, the self-attention mechanism was used to capture the items' long-term dependencies in sequential behaviors (Section III.A). Next, we designed a transaction graph that included item connections in chronological purchasing order. We divided the edges into incoming and outgoing edges to learn item interactions, reflecting changes in users' purchasing preferences. Using the transaction graph, we generated a correlation-specific item translation vector γ_r as relationships between entities, which includes sophisticated interactions between items in users' item sequences through GGNN (Section III.B). Finally, we optimized the metric function to score the interactions with the target item translation vector γ_j represented as the tail entity. In this module, we additionally created comprehensive item vectors γ_u that explicitly aggregated the item's information related to user-to-item and item-to-item interactions. Inspired by FT embedding, our distance function considered the direction of translation vectors to release the strict translational principle $h + r \approx t$. Owing to the flexible metric in the proposed method, we additionally considered the relationships between consumed items and target item vectors from the two perspectives with γ_u and $\gamma_{u'}$. Owing to the three generated vectors of user behavior-based item vector, comprehensive item vector, and item correlation vector as γ_u , $\gamma_{u'}$ and γ_r , respectively, we were able to optimize the translational embedding model $(\gamma_{u'} + \gamma_r)^T \gamma_j + (\gamma_j + \gamma_r)^T \gamma_u$ to find the next item (Section III.C).

A. User-Specific Item Translation Vector Embedding

Let U and I represent the user and item sets, where $u \in U$ denotes a user and $i \in I$ denotes an item. For each user u , we extracted every L successive items as a user action sequence. In this module, we generated a user-specific item translation vector $\gamma_u \in R^{L \times d}$. By reflecting users' long-term preferences in γ_u , AFTRec considers item transitions and users' purchasing history to model relationships between items using translation-based approaches.

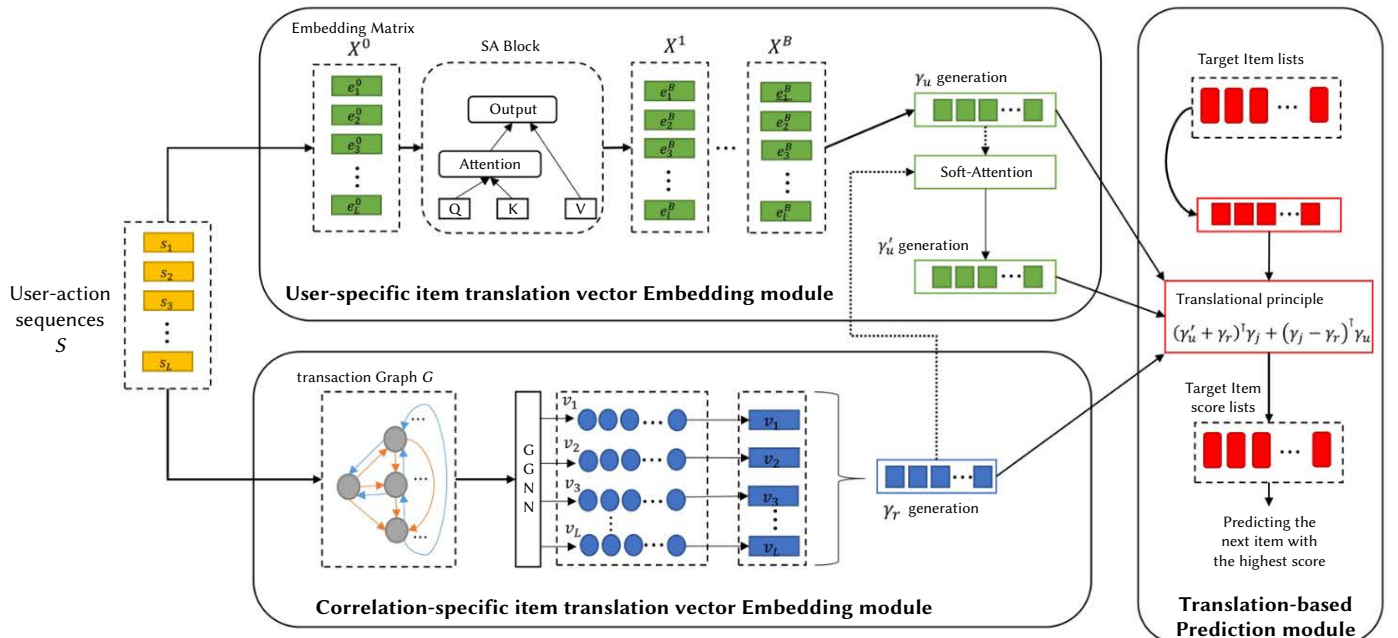


Fig. 4. Architecture of our proposed method.

Let $S_u = \{i_1, i_2, \dots, i_L\}$ denote the set of all L items ordered by timestamp. In this section, a user-specific item translation vector is generated for each previous item in $\{i_1, i_2, \dots, i_L\}$ as shown in Fig. 4. With the translational principle, AFTRec predicted the item for the $t + 1$ step-based translational distance with γ_u corresponding to the purchased item in step t ($0 < t < L$) in chronological order.

In S_u , we created an item embedding matrix $M \in R^{L \times d}$, where d is the latent dimensionality. In addition, we generated a learnable position embedding matrix $P \in R^{L \times d}$ as the purchasing order information in the user sequence. We obtained the item embedding lookup matrix $E \in R^{L \times d}$ by calculating $E = M + P$. To efficiently represent item translation vectors reflecting user preferences, we utilized stacked self-attention blocks (SABs) for E . The SAB consists of a multi-head attention and a pointwise feed-forward (FF) layer. Multi-head attention runs a scaled dot-product attention mechanism several times in parallel. Because it concatenates different representations of an item's dependencies from various perspectives, it is beneficial to consider multiple relationships jointly through a separate analysis. MHA was calculated as follows:

$$MHA(Q, K, V) = \text{Concat}[hd_1, hd_2, \dots, hd_h]W^{MHA} \quad (1)$$

$$hd_i = \text{Attention}(EW_i^Q, EW_i^K, EW_i^V) \quad (2)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (3)$$

where Q, K , and V denote the sets of queries, keys, and values, respectively. In addition, W^{MHA}, W_i^Q, W_i^K , and $W_i^V \in R^{d \times d}$ are learnable parameters, and \sqrt{d} is a scale factor that scales the dot products to avoid the vanishing gradient problem. We provided the item embedding lookup matrix E as input, which can be defined as a linear transformation of $Q = EW^Q, K = EW^K$, and $V = EW^V$. To reflect a realistic sequential behavior of a user to a user-specific item translation vector, we considered t items when generating the t -th purchased item translation vector. Therefore, we masked the queries and keys from $t+1$ to the last item. Then, the pointwise FF layer was calculated as follows:

$$SAB(X) = FFN\left(MHA(EW_i^Q, EW_i^K, EW_i^V)\right) \quad (4)$$

$$FFN(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \quad (5)$$

where $W_1, W_2 \in R^{d \times d}$ are the learnable parameters. In addition, $b_1, b_2 \in R^{1 \times d}$ are the bias parameters. A pointwise FF layer was applied to each item position separately to aggregate and normalize the attention outputs. Similar to [22], the pointwise FF layer included a residual connection and layer normalization, which are omitted in (4) for brevity. To efficiently improve the performance of capturing item transitions in long-term sequences, we stacked the SABs. The B -th ($B > 1$) block is defined as follows:

$$X^{(B)} = SAB(X^{(B-1)}) \quad (6)$$

where $X^{(0)} = E$. In this module, we obtained the output item embedding matrix $X' \in R^{d \times d}$ using stacked SABs. We then defined X' as the head γ_u in the transition space. In contrast to TransRec and MoHR, where the translation vectors are represented only by item embeddings, our proposed method is able to represent not only long-term preferences but also item transitions.

B. Correlation-Specific Item Translation Vector Embedding

In this section, we generate a relation vector that translates the interaction between the previous and next items for personalized recommendation in the same space. Therefore, we encoded complex item-to-item interactions in users' purchasing behaviors to correlation-specific item vectors $\gamma_r \in R^{L \times d}$ based on users' item sequences, as shown in Fig. 4.

Because the basic idea of graph neural networks (GNNs) [42] is to generate node embedding by aggregating the features and topological information from the neighbors, it ensures that GNNs are capable of efficiently capturing the interactions between nodes on graph-structured data. GGNN extends GNNs to sequential data, using a GRU as an update function to propagate information. Owing to the use of a GRU, GGNN selectively aggregates information of the neighbors, and thus, it is able to reduce the computational limitations and achieve a better performance. In this study, we converted personal item sequences to graph-structured data and learned the general relationships between consumed items in the e-commerce platform through GGNN, as described below.

1. Constructing a Session Graph

For user u , given the behavior sequence S_u , we designed a transaction graph G . Let $G_u = (V, E)$ be a directed graph, where each node denotes a purchased item at time t as $v_t \in I$, and each edge $(v_{t-1}, v_t) \in E$ denotes each link for a chronologically ordered pair of items. To represent chronological item-to-item relationships, we built an adjacency matrix using a sliding window that moved a unit distance ahead. For the user sequence S_u , we moved the window in a unit time and connected the links between neighboring item nodes.

An example of the construction of an adjacency matrix is shown in Fig. 5. The adjacency matrix $A \in R^{L \times 2L}$ is represented by two adjacency matrices $A^{BF}, A^{AF} \in R^{L \times L}$, which represent connections of earlier or later purchased items as incoming or outgoing edges in the transaction graph, respectively. All edges have normalized weights with connections between earlier or later items for each item. In Fig. 5, each graph representation of a user process through the adjacency matrix is based on item sequences of each user, and the transaction graph is generated by repeating this process for all users in the data-sparse environment.

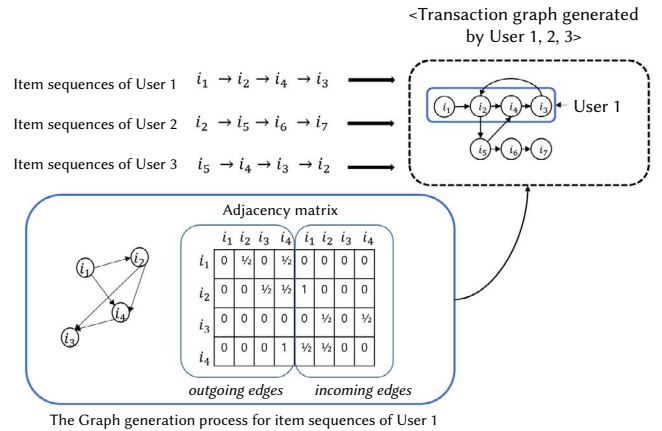


Fig. 5. An example of transaction graph generation process. Transaction graph is generated by the purchase history of users. The adjacency matrix is represented as a concatenation of two adjacency matrices, which link earlier or later purchased items, respectively.

2. Item-to-Item Interaction Learning

After constructing a transaction graph G_u of each user, we adopted a GGNN to learn item-to-item relations. Owing to the gating mechanism of a GRU, a GGNN can tackle the vanishing gradient and computational limitations by selectively gathering information from the other nodes to update the hidden state of each node. Let $h_i \in R^d$ denote the embedded node vector of the corresponding item v_i and H denote the set of all item node vectors. For the initialization step, the aggregation information a_i is defined as the concatenation of two types of adjacency matrices $A_i^{BF}, A_i^{AF} \in R^{1 \times d}$ corresponding to the target node h_i :

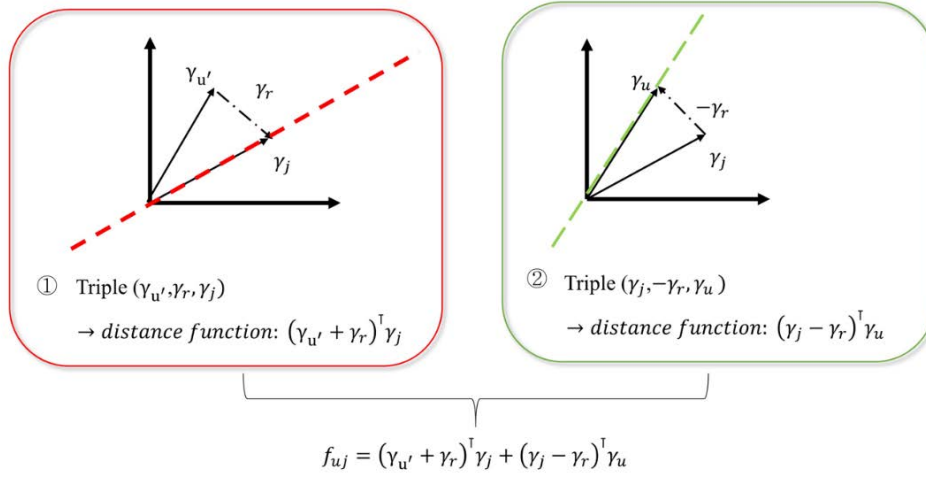


Fig. 6. Illustrations of translational principle for TransRec and AFTRec.

$$a_{i,BF}^t = A^{BF}([h_1^{t-1}, \dots, h_L^{t-1}]W_{BF}) + b_{BF} \quad (7)$$

$$a_{i,AF}^t = A^{AF}([h_1^{t-1}, \dots, h_L^{t-1}]W_{AF}) + b_{AF} \quad (8)$$

$$a_i^t = [a_{i,BF}^t; a_{i,AF}^t] \quad (9)$$

where $W_{BF}, W_{AF} \in R^{d \times d}$ are learnable parameters; $b_{BF}, b_{AF} \in R^d$ are bias parameters; $[h_1^{t-1}, \dots, h_L^{t-1}]$ is the list of item node states; and $[\cdot; \cdot]$ is the concatenation operation.

Then, the computation steps of updating h_i are defined as follows:

$$z_i^t = \sigma(W_z a_i^t + U_z h_i^t) \quad (10)$$

$$r_i^t = \sigma(W_r a_i^t + U_r h_i^t) \quad (11)$$

$$\tilde{h}_i^t = \tanh(W_o a_i^t + U_o (r_i^t \odot h_i^{t-1})) \quad (12)$$

$$h_i^t = (1 - z_i^t) \odot h_i^{t-1} + z_i^t \odot \tilde{h}_i^t \quad (13)$$

where $W_z, W_r, W_o \in R^{2d \times d}$, $U_z, U_r, U_o \in R^{d \times d}$ are learnable parameters. In addition, z_i^t and r_i^t are the update and reset gates, respectively. The reset gate determines the amount of past information that must be preserved or discarded. The update gate determines the amount of past information that must be passed along to the future. Moreover, σ denotes the logistic sigmoid function, and \odot denotes the element-wise multiplication. This procedure was computed in a manner similar to the GRU. After this procedure, the corresponding items of all updated nodes were defined as the relations γ_r that contain high-level item-to-item interactions and short-term user interests in the transaction graph.

C. Optimization and Target Item Prediction

After obtaining the user and item translation vectors as the head and relation, respectively, we could predict the target item as the tail by optimizing the translational embedding model. In previous translation-based recommendations, (h, r, t) was modeled by the same translational principle $h + r \approx t$ in KG embedding techniques (e.g., TransE and TransR [43]). However, the translational principle $h + r \approx t$ is too strict to model the complex and diverse interactions between entities and relations (e.g., symmetric/transitive/one-to-many/many-to-one/many-to-many relations). To consider an item's diverse information related to personal preferences in a metric space, we extended the FT to generate flexible translation vectors with respect to multiple entities and relations. Originally, FT embedded multiple entities and relations by optimizing $(h + r)^\top t + (t - r)^\top h$. Given an ideal embedding $h + r \approx t$, FT applies $h + r \approx \rho t, \rho > 0$ by considering directions of vectors $h + r$ and t . To balance the constraints on the head and tail during training, FT considers both directions of vectors

t and $h + r$ and h and $t - r$. Thus, it can flexibly capture more diverse and complex relationships between the head and tail.

For each triple (h, r, t) , we can create an inverse triple (t, r^{-1}, h) , which has also been used in [44], [45]. Thus, we can convert the translational principle $h + r \approx t$ to $t - r \approx h$. Using the FT principle, we can also apply $t - r \approx \rho h, \rho > 0$. For personal recommendations, we treat two user-specific item vectors as head entities: the user behavior-based item translation vector γ_u and the attentive item translation vector $\gamma_{u'}$. An attentive item translation vector strengthens the crucial information in the relationship between general purchased items and user preferences. Therefore, we applied the soft-attention mechanism [46] for long-term user interest and sophisticated item relations and then successfully aggregated the context pairs of user interest-to-item relations. The attentive item translation vector $\gamma_{u'}$ is defined as

$$m = \tanh(\text{ReLU}(W_3(W_4 \gamma_u + W_5 \gamma_r) + b_3)) \quad (14)$$

$$\alpha = \frac{\exp(m)}{\sum_{l=1}^L \exp(m_l)} \quad (15)$$

$$\gamma_{u'} = \alpha \gamma_u \quad (16)$$

where $W_3, W_4, W_5 \in R^{d \times d}$ are learnable parameters, and $b_3 \in R^d$ is the bias parameter. By considering the heads $\gamma_{u'}, \gamma_u$ and the relation γ_r , we can seek the tail γ_t to predict a suitable item for the user's dynamic preferences. Fig. 6 illustrates translational embedding models of TransRec and our AFTRec. In our translational embedding model, we consider the directions of the vectors $(\gamma_{u'} + \gamma_r)$ with γ_j , and $(\gamma_j - \gamma_r)$ with γ_u .

Using balanced learning for the interactions of two triple sets $(\gamma_{u'}, \gamma_r, \gamma_j)$ and $(\gamma_{u'}, \gamma_r, \gamma_j)$ in a translation space, as shown in Fig. 6, AFTRec can flexibly capture diverse user and target item relations using different perspectives for the personal preferences of users. Finally, the model scores can be formulated as follows:

$$f_{uj} = (\gamma_{u'} + \gamma_r)^\top \gamma_j + (\gamma_j - \gamma_r)^\top \gamma_u \quad (17)$$

Based on our model's score, as shown in Fig. 6, AFTRec aims to maximize the probability of a true item under relationships in a user's behavior sequence. We adopted the binary cross-entropy loss for the optimization of the translation-based methods proposed by [47]–[49]. Given positive item set I and negative item set I' , positive item $j \in I$ and negative item $j' \in I'$ are uniformly sampled. Then, we optimize the loss function as follows:

$$\mathcal{L} = \sum_{u \in U} \sum_{j \in I} -[\log \sigma(f_{uj}) + \sum_{j' \in I'} \log(1 - \sigma(f_{uj'}))] \quad (18)$$

where σ is the logistic sigmoid function used to obtain the predicted probability of a triple. In this model, we updated the parameters using an Adam optimizer [50] and regularized the parameters based on L_2 regularization to prevent overfitting. In the training process, for items purchased before the last purchased item, AFTRec modeled item vectors to predict the next item using the previous item based on our translational principle. AFTRec finally recommended an appropriate item for the user with the highest f_{ij} score with the user- and correlation-specific item translation vector for the latest purchased item.

IV. EXPERIMENTS

A. Datasets

We evaluated AFTRec on five public datasets for real-world applications. All datasets had diverse domains and sizes. The statistics of all datasets are reported in Table I. For comparison with translation-based models that require standardized relationships between users and items, we used datasets from the Amazon and Steam platforms, which define specific relationship types between user-to-user and item-to-item pairs. We take five domains: “Beauty,” “Toys and games (Toys),” “Clothing, shoes, and jewelry (Clothing),” and “Automotive” from Amazon review datasets in [51], and “Games” from Steam datasets generated in [33]. Amazon datasets were used as sparse datasets, whereas the Steam dataset was used as dense dataset. In this section, we demonstrate our performance for sparse datasets using Amazon datasets, and we experiment with our recommendation performance on dense datasets using the Steam dataset. All the datasets contain various user-to-item interaction data (e.g., user ratings and reviews). We followed the methods used by Kang and McAuley [33], and Wu et al. [34] to preprocess datasets to sort items in the sequential order of user sequences. First, we ordered the review behaviors as positive feedbacks by the timestamps. Second, we discarded users with fewer than five related-item interactions. Then, we transformed the users’ review data to become a sequential dataset indicating the order of each user’s purchase items.

For each user, we split the user’s historical sequences S_u into three parts, as done by Kang and McAuley [33], and Wu et al. [34]: (1) the most recent interaction in S_u as the testing set, (2) the next interaction as the validation set, and (3) the remaining interactions as the training set.

B. Evaluation Metric

We used two common Top-K recommendations: the hit rate (HR@10) and normalized discounted cumulative gain (nDCG@10). Here, HR@10 is the rate of positive items in the top-10 recommended items, and nDCG@10 is a ranking measurement for the positions of the positive items in the top-10 recommended items. For the computational cost, we followed the previous mentioned works [33], [34]. We randomly sampled 100 negative and 1 positive item for each user and ranked them for evaluation.

C. Comparison Methods

To evaluate the performance of AFTRec, we compared it with the following eight competitive baselines:

POP: Simple baseline recommendation model that recommends the most popular items in the training set.

CML: CF-based method that applies metric learning instead of MF. It learns a metric to minimize similar user and item pairs.

FPMC: Sequential RS that combines MF and factorized first-order MC. It captures long-term user interests and item-to-item transitions by utilizing the characteristics of both methods. TransRec: Baseline translation-based method for sequential recommendations. It embeds

users and items into the transition space and models three-component relationships between a user, previously visited items, and target item.

MoHR: Translation-based method that minimizes the distance between relevant item pairs in the translation space. It exhibits different relation types (e.g., also-viewed/also-bought) between user and item pairs and is integrated into the translational embedding model.

SASRec: Self-attention-based sequential recommendation model inspired by a transformer in NLP. It captures the long-term user interest in predicting the next item through multiple stacked SABs.

TiSASRec: Self-attention-based sequential recommendation model. Unlike SASRec, which considers the absolute time position of items, TiSASRec uses relative time intervals for positioning the encodings of items in stacked SABs.

D. Implementation Details

During the experiments, we implemented AFTRec using the Adam optimizer with momentum exponential decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We set the batch size to 128 and the maximum sequence length to 50 for all datasets. In AFTRec, we set the number of SABs to two and used single-head self-attention layers to generate the user translation vector. We set the number of links in the transaction graph to three for learning the item relations. For comparison with competitive baselines, the hyperparameters were tuned through a grid search. The learning rate was $\{0.1, 0.001, 0.0001, 0.00001\}$, and the dropout rate was $\{0.2, 0.5\}$. For SASRec and TiSASRec, we set the number of SABs to two and used single-head self-attention layers. For SASRec and TiSASRec, the embedding dimensions were set to 50. For TransRec and MoHR, the embedding dimensions were set to 10. Except for POP, CML, FPMC, and TransRec, the batch size was set to 128. For SASRec and TiSASRec, the maximum sequence lengths were 50. We set all other parameters according to the respective baseline papers.

E. Recommendation Performance

Tables II and III show a performance comparison of sequential recommendations and translation-based recommendations with HR@10 and nDCG@10 on four sparse datasets and one dense dataset. On sparse datasets, AFTRec achieved the best performance for both the HR@10 and nDCG@10 metrics. These results show that AFTRec outperforms sequential recommendations using only the self-attention mechanism and translation-based sequential recommendations to resolve the data sparsity problem in the data-sparse environment such as e-commerce recommendation. Several observations of the competitive baselines are shown in Table II. For the Beauty and Toys datasets, POP, which is a traditional recommendation, achieves the worst performance in terms of nDCG and HR. TiSASRec achieved the second-best performance among the baseline methods in terms of nDCG and HR on the Beauty, Toys, and Clothing datasets. In addition, SASRec achieved the second-best performance in nDCG and HR among the baselines on the Beauty dataset.

The proposed model showed better nDCG@10 performance than the existing model for all datasets and better HR@10 performance than the existing models on sparse datasets (Table III). In particular, the proposed model showed the greatest improvement in nDCG and HR performance compared to the existing embedding-based recommendation model for the Clothing dataset. For all datasets, CML, which applies a metric function instead of MF, achieved the worst performance in terms of nDCG and HR. For sparse datasets, MoHR achieved the second-best performance in terms of nDCG and HR.

Compared with these baselines, the proposed AFTRec achieved the best performance on the four datasets. This is because our method represents the user’s short-term and long-term interests as user translation vectors through self-attention to user sequences and

TABLE I. STATISTICS OF DATASETS USED IN EVALUATIONS

Dataset	# Users	# Items	# Actions	Avg of actions/user
Automotive	34,315	40,287	183,567	5.35
Beauty	52,204	57,289	394,908	7.56
Clothing	184,050	174,484	1,068,972	5.81
Toys	57,617	69,147	410,920	7.39
Steam	335,730	13,047	4,213,117	12.59

TABLE II. COMPARISON OF RECOMMENDATION PERFORMANCE ON FIVE PUBLIC DATASETS AND FOUR SEQUENTIAL RECOMMENDATIONS. THE BEST PERFORMING METHOD IS IN BOLDFACE. THE LATENT DIMENSION SIZE D FOR ALL BASELINES WAS SET TO 50

Dataset	Metric	PopRec	FPMC	SASRec	Ti-SASRec	AFTRec
Automotive	nDCG@10	0.2084	0.1981	0.2288	0.2509	0.4875
	HR@10	0.3481	0.3210	0.3716	0.4032	0.8992
Beauty	nDCG@10	0.2277	0.2532	0.3211	0.3126	0.4325
	HR@10	0.4003	0.4070	0.4852	0.4734	0.8571
Clothing	nDCG@10	0.2166	0.2076	0.2214	0.2445	0.4667
	HR@10	0.3661	0.3478	0.3853	0.3974	0.8872
Toys	nDCG@10	0.2048	0.2651	0.3136	0.3177	0.4730
	HR@10	0.3601	0.4170	0.4663	0.4920	0.8596
Steam	nDCG@10	0.4728	0.5297	0.6211	0.6228	0.5716
	HR@10	0.7297	0.7830	0.8716	0.8657	0.9036

TABLE III. COMPARISON OF RECOMMENDATION PERFORMANCE ON FIVE PUBLIC DATASETS AND THREE TRANSLATION-BASED SEQUENTIAL RECOMMENDATIONS. THE BEST PERFORMING METHOD IS IN BOLDFACE. THE LATENT DIMENSION SIZE D FOR ALL BASELINES WAS SET TO 10

Dataset	Metric	CML	TransRec	MoHR	AFTRec
Automotive	nDCG@10	0.1793	0.2034	0.3478	0.3845
	HR@10	0.3062	0.3332	0.5382	0.7260
Beauty	nDCG@10	0.2532	0.2666	0.3635	0.4004
	HR@10	0.4070	0.4125	0.5550	0.7416
Clothing	nDCG@10	0.1904	0.2111	0.3015	0.4457
	HR@10	0.3307	0.3608	0.4919	0.7024
Toys	nDCG@10	0.2437	0.2890	0.4151	0.4185
	HR@10	0.4015	0.4474	0.6061	0.7734
Steam	nDCG@10	0.4699	0.5287	0.5598	0.5835
	HR@10	0.7481	0.7842	0.7983	0.7020

high-level item relations as item translation vectors. By mapping user- and item-specific vectors onto the head and the relation into the transition space, we can utilize the advantages of self-attention-based methods and translation principles. In addition, we modeled the interactions between the user and target item efficiently by optimizing the translational embedding model, which considers the directions of both user and target item vectors toward a FT. This shows that the modeling of translational relationships with users, items, and heterogeneous items is generally efficient in capturing a user’s long-term interest and short-term item transitions by leveraging a translation function for the given user-to-item interactions. Neural-network-based sequential recommendations are generally superior for predicting personal recommendations on relatively large datasets with respect to interactions between users and items. In contrast, on relatively small datasets with respect to interactions between users and items, translation-based sequential recommendations can provide better recommendations by utilizing interactions between user and item translation vectors captured by transitional principle-based KG embedding techniques.

F. Limitations for AFTRec

Tables II and III show a performance comparison of sequential recommendations and translation-based recommendations with HR@10 and nDCG@10 on four sparse datasets and one dense dataset.

For the Steam dataset, which is a dense dataset (Table II), TiSASRec achieved the best performance in terms of nDCG. In contrast, AFTRec achieved the best performance in terms of HT. For nDCG, self-attention-based models were advantageous for predicting the next item for dense datasets in sequential recommendations. However, AFTRec applies a self-attention mechanism to generate user-specific item translation vectors. Therefore, a user’s item preferences with self-attention are advantageous for showing candidate items that include true items in terms of HR. Because the proposed model comprehensively learns the user’s purchase characteristics and the comprehensive correlations between the users’ purchased items, the nDCG performance is slightly lowered, but our model shows better HR performance, indicating whether the true item is exposed to the recommendation candidates. Among the sparse datasets, the proposed model showed the greatest performance improvement on the Automotive dataset, which is a representative sparse dataset, and the experimental results show that the proposed model has better recommendation performance than the existing models.

On the Steam dataset, MoHR achieved the best performance in terms of nDCG with dimensions of 10 (Table III). Considering that the user-specific item translation vector and correlation-specific item translation vector generated by the proposed model are trained by a neural network, the experimental results show that the recommendation performance of the proposed model is slightly lower

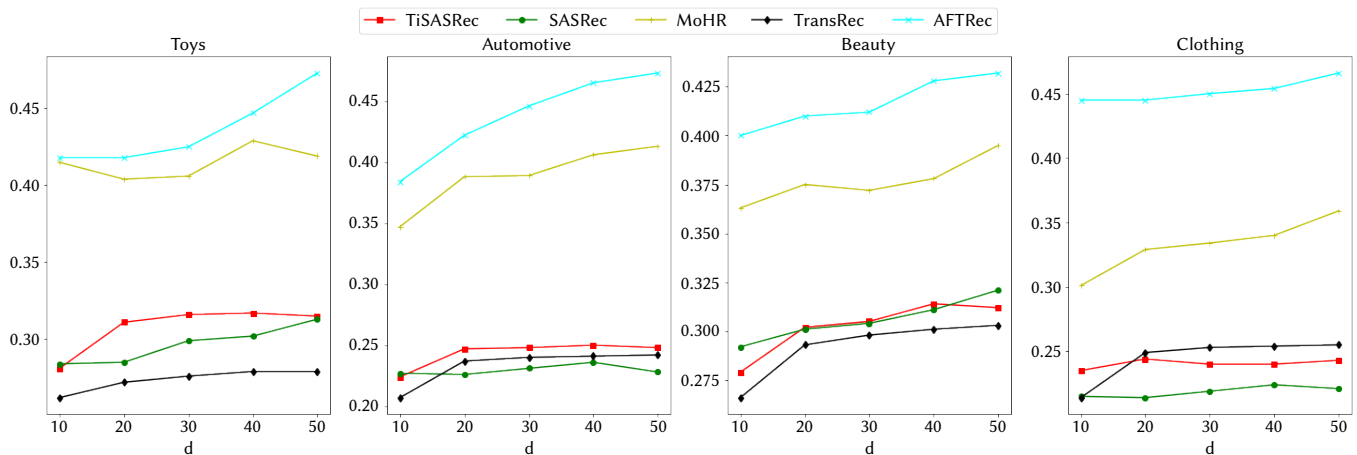


Fig. 7. Comparison of recommendation performance on four datasets (nDCG) with varying latent dimension size d of 10 to 50.

TABLE IV. COMPARISON OF RECOMMENDATION PERFORMANCE ON FOUR DATASETS (NDCG) WHEN VARYING THE NUMBER OF SELF-ATTENTION BLOCKS (SABs) OF 1 TO 3

Dataset	Metric	Number of SABs		
		1	2	3
Automotive	nDCG@10	0.4371	0.4875	0.4105
Beauty	nDCG@10	0.4073	0.4325	0.4264
Clothing	nDCG@10	0.4016	0.4667	0.4090
Toys	nDCG@10	0.3819	0.4730	0.4083
Steam	nDCG@10	0.5104	0.5716	0.5367

than that of MoHR when the size of the data dimension is 10. However, because the proposed model learned various features extracted from purchased items as item translation vectors, it showed a higher recommendation performance than the existing recommendation models, which do not properly reflect the sequential purchase characteristics of the user.

G. Hyperparameter Study

We conducted an additional experiment that varied the dimension size on four sparse datasets to investigate the performance changes based on different embedding dimension sizes of d . The dimension size affects the item embedding size of self-attention and GGNN for entities and relations. We changed the dimension sizes from $\{10, 20, 30, 40, 50\}$, and the nDCG@10 results are shown in Fig. 7. For the Amazon datasets, our model outperformed the baselines. From Fig. 7, TransRec, MoHR, and AFTRec achieved better performance as the latent dimension $d \geq 30$ increased on the Automotive, Beauty, and Clothing datasets. By contrast, for the Toys dataset, the performance of the MoHR peaked when $d = 40$. A dimension size of 40 represents sufficient information for MoHR on the Toys dataset. In addition, translation-based models generally have more advanced performance than neural network-based models, such as SASRec and TiSASRec, on sparse datasets.

In Fig. 7, SASRec and TiSASRec show the following aspects. The performance of TiSASRec peaks when $d = 40$ on the Toys and Automotive datasets. For Beauty and Clothing datasets, TiSASRec achieved better performance as d increased. In addition, for the Toys and Beauty datasets, SASRec for Automotive and Clothing datasets, the performance of SASRec peaked when $d = 40$. It is indicated that a dimension size of 40 provides sufficient information for SASRec on Automotive and Clothing datasets. Thus, we find that the dimension size d affects the model's ability to represent sufficient information for user preferences.

We also changed the number of SABs to efficiently learn more complex global preferences of users (Table IV). For all datasets,

AFTRec exhibited the best performance on nDCG@10 when using two SABs. The performance of AFTRec increased until the number of SABs was set to two, but AFTRec decreased performance with more than two SABs. From these results, we found that AFTRec has a more stable performance with two SABs.

V. CONCLUSION

In this study, we proposed AFTRec, a novel translation-based sequential recommendation method for sequential personal historical behaviors for the data sparsity problem. The process maps user preferences and sophisticated item relations to embedding vectors to model the interactions between users and items using the transitional principle. The proposed method includes three main processes. First, for the user-specific item translation vector, we utilized SABs to adaptively capture short- and long-term user preferences in user historical sequences. Second, we designed a transaction graph that links relevant items in terms of timestamps. We applied a GGNN to the transaction graph to generate the item vector, which represents complex interactions between chronologically relevant items and embeds a correlation-specific item translation vector for each item. Third, we employed an attentive user vector using a soft-attention mechanism to jointly learn user-to-item relations in diverse forms of user embedding. After considering the item translation vectors as the heads and the relation vectors, AFTRec models the interactions between the user and items in the same translation space. Because our translational embedding model considers the direction of the embedding vectors, it flexibly provides suitable recommendations for user preferences.

We conducted experiments to evaluate our method on the Automotive, Clothing, Beauty, and Toys datasets collected by the Amazon platform and the Game dataset collected by the Steam platform. The experimental results demonstrate that our method outperforms state-of-the-art baselines in terms of both nDCG and HR on a sparse dataset. Therefore, the experimental results demonstrate

that our model is appropriate for predicting the next item in sparse datasets. In the future, we plan to improve the performance of our model and extend it by incorporating complex context-level user information, such as user groups, locations, and devices.

ACKNOWLEDGMENT

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the Innovative Human Resource Development for Local Intellectualization support program (IITP-2022-RS-2022-00156360) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation).

REFERENCES

- [1] C. Feng, J. Liang, P. Song, and Z. Wang, "A fusion collaborative filtering method on sparse data in recommender systems," *Information Sciences*, vol. 421, pp. 365-379, 2020, doi: 10.1016/j.ins.2020.02.052.
- [2] J. Bobadilla, S. Alonso, and A. Hernando, "Deep learning architecture for collaborative filtering recommender systems," *Applied Sciences*, vol. 10, no. 7, pp. 2441, 2020, doi:10.3390/app10072441.
- [3] A. Gazdar, and L. Hidri, "A new similarity measure for collaborative filtering based recommender systems," *Knowledge-Based Systems*, vol. 188, 2020, doi: 10.1016/j.knsys.2019.105058.
- [4] Y. Koren, R. Bell, and Chris Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8 pp. 30-37, 2009, doi: 10.1109/MC.2009.263.
- [5] D. Jannach, M. Ludewig, and L. Lerche, "Session-based item recommendation in e-commerce: on short-term intents, reminders, trends and discounts," *User Modeling and User-Adapted Interaction*, vol. 27, pp.351-392, 2017, doi: 10.1007/s11257-017-9194-1.
- [6] A. Luo, P. Zhao, Y. Liu, F. Zhuang, D. Wang, J. Xu, et al., "Collaborative Self-Attention Network for Session-based Recommendation," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Yokohama, Japan, 2020, pp.2591-2597.
- [7] S. Sun, Y. Tang, Z. Dai, and F. Zhou, "Self-Attention Network for Session-Based Recommendation with Streaming Data Input," *IEEE Access*, vol. 7, pp. 110499-110509, 2019, doi: 10.1109/ACCESS.2019.2931945.
- [8] D. Hu, L. Wei, W. Zhou, X. Huai, Z. Fang, and S. Hu, "PEN4Rec: Preference Evolution Networks for Session-based Recommendation," *arXiv preprint arXiv:2106.09306*, 2021.
- [9] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme, "Factorizing Personalized Markov Chains for Next-basket Recommendation," in *Proceedings of the 19th International Conference on World Wide Web*, North Carolina, USA, 2010, pp. 811-820.
- [10] D. W. Hogg, and D. Foreman-Mackey, "Data analysis recipes: Using markov chain monte carlo," *The Astrophysical Journal Supplement Series*, vol. 236, no.1, pp.11, 2018, doi: 10.3847/1538-4365/aab76e.
- [11] P. Tengkiattrakul, S. Maneeroj, and A. Takasu, "Attentive Hybrid Collaborative Filtering for Rating Conversion in Recommender Systems," in *Proceedings of the International Conference on Energy, Water and Environment*, Venice, Italy, 2021, pp.151-165.
- [12] Y. Zhang, Y. He, J. Wang, and J. Caverlee, "Adaptive Hierarchical Translation-based Sequential Recommendation," in *Proceedings of the Web Conference 2020*, Taipei, Taiwan, 2020, pp. 2984-2990.
- [13] Y. Ding, Y. Ma, W. Wong, and T. S. Chua, "Modeling Instant User Intent and Content-level Transition for Sequential Fashion Recommendation," *IEEE Transactions on Multimedia*, preprint.
- [14] A. Garcia-Duran, R. Gonzalez, D. Onoro-Rubio, M. Niepert, and H. Li, "Transrev: Modeling reviews as translations from users to items," *Advances in Information Retrieval*, vol.12035, pp.234-248, 2020, doi: 10.1007/978-3-030-45439-5_16.
- [15] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, 2017, pp. 2724-2743, DOI: 10.1109/TKDE.2017.2754499.
- [16] M. Brandalero, M. Shafique, L. Carro, and A. C. S. Beck, "Transrec: Improving adaptability in single-ISA heterogeneous systems with transparent and reconfigurable acceleration," in *2019 Design, Automation & Test in Europe Conference & Exhibition*, Florence, Italy, 2019, pp. 582-585.
- [17] Y. Tay, L. Anh-Tuan, and S. C. Hui, "Latent relational metric learning via memory-based attention for collaborative ranking," in *Proceedings of the 2018 World Wide Web Conference*, Lyon, France, 2018, pp. 729-739.
- [18] C. K. Hsieh, L. Yang, Y. Cui, T. Y. Lin, S. Belongie, and D. Estrin, "Collaborative metric learning," in *Proceedings of the 26th international conference on world wide web*, Geneva, Switzerland, 2017, pp. 193-201.
- [19] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multirelational data," in *Advances in Neural Information Processing Systems*, Lake Tahoe, Nevada, 2013, pp. 2787- 2795.
- [20] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," *arXiv preprint arXiv:1511.05493*, 2015.
- [21] J. Feng, M. Huang, M. Wang, M. Zhou, Y. Hao and X. Zhu, "Knowledge graph embedding by flexible translation," in *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*, Cape Town, South Africa, 2017.
- [22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., "Attention is all you need," In *Advances in neural information processing systems*, California, USA, 2017, pp. 5998-6008.
- [23] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.
- [24] S. Kabbur, X. Ning, and G. Karypis, "Fism: factored item similarity models for top-n recommender systems," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, Chicago, USA, 2013, pp. 659-667.
- [25] X. He, Z. He, J. Song, Z. Liu, Y. G. Jiang, and T. S. Chua, "Nais: Neural attentive item similarity model for recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 12, 2018, pp. 2354-2366, doi: 10.1109/TKDE.2018.2831682.
- [26] J. Bobadilla, A. Gutiérrez, S. Alonso and A. González-Prieto, "Neural Collaborative Filtering Classification Model to Obtain Prediction Reliabilities," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, 2022, pp. 18-26, doi: 10.9781/ijimai.2021.08.010.
- [27] J. Bobadilla, R. Lara-Cabrera, A. González-Prieto and F. Ortega, "DeepFair: Deep Learning for Improving Fairness in Recommender Systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, 2021, pp. 86-94, doi: 10.9781/ijimai.2021.11.001.
- [28] B. Hidasi, and A. Karatzoglou, "Recurrent neural networks with top-k gains for session-based recommendations," in *Proceedings of the Conference on Information and Knowledge Management*, Turin, Italy, 2018, pp. 843-852.
- [29] P. M. Gabriel De Souza, D. Jannach, and A. M. Da Cunha, "Contextual hybrid session-based news recommendation with recurrent neural networks," *IEEE Access*, vol. 7, 2019, pp.169185-169203, doi: 10.1109/ACCESS.2019.2954957.
- [30] S. Sun, Y. Tang, Z. Dai, and F. Zhou, "Self-attention network for session-based recommendation with streaming data input," *IEEE Access*, vol. 7, 2019, pp. 110499-110599, doi: 10.1109/ACCESS.2019.2931945.
- [31] C. Xu, J. Feng, P. Zhao, F. Zhuang, D. Wang, Y. Liu, et al., "Long- and short-term self-attention network for sequential recommendation," *Neurocomputing*, vol. 423, 2021, pp. 580-589, doi: 10.1016/j.neucom.2020.10.066.
- [32] J. Li, P. Ren, Z. Chen, Z. Ren, T. Lian, and J. Ma, "Neural attentive session-based recommendation," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Singapore, 2017, pp. 1419-1428.
- [33] W. C. Kang, and J. McAuley, "Self-attentive sequential recommendation," in *2018 IEEE International Conference on Data Mining*, Singapore, 2018, pp. 197-206.
- [34] L. Wu, S. Li, C. J. Hsieh, and J. Sharpnack, "SSE-PT: Sequential recommendation via personalized transformer," in *Fourteenth ACM Conference on Recommender Systems*, Brazil, 2020, pp. 328-337.
- [35] J. Li, Y. Wang, and J. McAuley, "Time interval aware self-attention for sequential recommendation," in *Proceedings of the 13th international conference on web search and data mining*, Texas, USA, 2020, pp. 322-330.
- [36] J. Tang, and K. Wang, "Personalized top-n sequential recommendation via convolutional sequence embedding," in *Proceedings of the ACM Conference on Web Search and Data Mining*, California, USA, 2018, pp. 565-573.

- [37] F. Yuan, A. Karatzoglou, I. Arapakis, J. M. Jose, and X. He, "A simple convolutional generative network for next item recommendation," in *Proceedings of the ACM Conference on Web Search and Data Mining*, Melbourne, Australia, 2019. pp. 582–590.
- [38] W. C. Kang, M. Wan, and J. McAuley, "Recommendation Through Mixtures of Heterogeneous Item Relationships," in *Proceedings of ACM Conference on Information and Knowledge Management*, Turin, Italy, 2018, pp. 1143-1152.
- [39] B. Wu, X. He, Z. Sun, L. Chen, and Y. Ye, "ATM: An attentive translation model for next-item recommendation," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, 2020, pp. 1448-1459, doi: 10.1109/TII.2019.2947174.
- [40] R. Pasricha, and J. McAuley, "Translation-based factorization machines for sequential recommendation," in *Proceedings of the 12th ACM Conference on Recommender Systems*, Vancouver, Canada, 2018, pp.63-71.
- [41] Y. Zhan, Y. He, J. Wang, and J. Caverlee, "Adoptive hierarchical translation-based sequential recommendation," in *Proceedings of the Web Conference 2020*, Taipei, Taiwan, 2020, pp. 2984-2990.
- [42] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE transactions on neural networks*, vol. 20, no. 1, 2009, pp.61-80, doi: 10.1109/TNN.2008.2005605.
- [43] C. Ma, P. Kang, and X. Liu, "Hierarchical gating networks for sequential recommendation," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, Alaska, USA, 2019, pp. 825-833.
- [44] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proceedings of 29th AAAI conference on artificial intelligence*, Texas, USA, 2015, pp. 2181–2187.
- [45] T. Lacroix, N. Usunier, and G. Obozinski, "Canonical tensor decomposition for knowledge base completion," in *Proceedings of the International Conference on Machine Learning*, Stockholm, Sweden, 2018, pp. 2863-2872.
- [46] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
- [47] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2d knowledge graph embeddings," in *Thirty-second AAAI conference on artificial intelligence*, Louisiana, USA, 2018.
- [48] Z. Sun, Z. H. Deng, J. Y. Nie, and J. Tang, "RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space," in *International Conference on Learning Representations*, Louisiana, USA, 2018.
- [49] S. M. Kazemi, and D. Poole, "Simple embedding for link prediction in knowledge graphs," in *The Thirty-second Annual Conference on Neural Information Processing Systems*, Montreal, Canada, 2018.
- [50] D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [51] J. J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Image-based recommendations on styles and substitutes," in *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, Santiago, Chile, 2015, pp. 43-52.



Min-Ji Seo

She received the B.S. in Computer Science and Engineering (2015) and the M.S. and ph.D in software convergence from Soongsil University, Seoul, Korea, in 2017 and 2022, respectively. Her current research interests are data leak prevention, speech emotion recognition and deep learning.



Myung-Ho Kim

He received the B.S. in Computer Engineering from Soongsil University, Seoul, Korea (1989) and the M.S. and ph.D in Computer Engineering from POSTECH, Pohang, Korea, in 1991 and 1995, respectively. Now he is a professor in department of software convergence from Soongsil University. His current research interests are deep learning and block chain and cloud computing.

Synthetic Aperture Radar Automatic Target Recognition Based on a Simple Attention Mechanism

Chiagoziem C. Ukwuoma¹, Qin Zhiguang^{1*}, Bole W. Tienin², Sophyani B. Yussif³, Chukwuebuka J. Ejiyi¹, Gilbert C. Urama³, Chibueze D. Ukwuoma⁴, Ijeoma A. Chikwendu²

¹ School of Information and Software Engineering, University of Electronic Science and Technology of China (China)

² School of Information and Communication Engineering, University of Electronic Science and Technology of China (China)

³ School of Computer Science and Engineering, University of Electronic Science and Technology of China (China)

⁴ Department of Physics- Electronics, Federal University of Technology Owerri (Nigeria)

Received 13 August 2021 | Accepted 10 January 2023 | Published 6 February 2023



ABSTRACT

A simple but effective channel attention module is proposed for Synthetic Aperture Radar (SAR) Automatic Target Recognition (ATR). The channel attention technique has shown recent success in improving Deep Convolutional Neural Networks (CNN). The resolution of SAR images does not surpass optical images thus information flow of SAR images becomes relatively poor when the network depth is raised blindly leading to a serious gradients explosion/vanishing. To resolve the issue of SAR image recognition efficiency and ambiguity trade-off, we proposed a simple Channel Attention module into the ResNet Architecture as our network backbone, which utilizes few parameters yet results in a performance gain. Our simple attention module, which follows the implementation of Efficient Channel Attention, shows that avoiding dimensionality reduction is essential for learning as well as an appropriate cross-channel interaction can preserve performance and decrease model complexity. We also explored the One Policy Learning Rate on the ResNet-50 architecture and compared it with the proposed attention based ResNet-50 architecture. A thorough analysis of the MSTAR Dataset demonstrates the efficacy of the suggested strategy over the most recent findings. With the Attention-based model and the One Policy Learning Rate-based architecture, we were able to obtain recognition rate of 100% and 99.8%, respectively.

KEYWORDS

Attention Mechanism, Automatic Target Recognition, Deep Convolutional Neural Network, Synthetic Aperture Radar.

DOI: 10.9781/ijimai.2023.02.004

I. INTRODUCTION

IMAGES of the Earth's surface taken by employing Synthetic Aperture Radar (SAR) systems, an observation tool, regardless of the weather condition, is referred to as SAR images. The SAR Automatic Target Recognition (ATR), which is an essential part of SAR image interpretation, is one long-term research complex problem for researchers across the globe since it is generally applied in not only the military field but also in the civilian ones mainly since it is usable in any weather and time of the day. Contrary to the optical images with colors considered rich, SAR images can be distinguished by the possession of solid grayscale pixels with regions that have high intensities representing the targets. SAR image classification, which tags per pixel in accordance with one or more retrieved characteristics, is crucial to SAR image comprehension. In a broad sense, SAR image analysis might be used widely in a variety of fields, including monitoring of the environment and natural resources [1], hydrological and agribusiness modeling [2], and urban planning [3]. The architecture

of SAR ATR which is basic is composed of three components which are detection and discrimination, alongside classification [4]. In the first component – detection, target regions or areas are extracted by a detector named Constant False Alarm Rate (CFAR) detector [5]. In the second component – discrimination, the application of the discriminator is for the identification of the candidate areas that are located by the targets with respect to the output of stage one. The third component – classification makes use of a classifier to identify the category of every target type.

Convolutional Neural networks (CNN) that are deep learning-based have been seen as one of the approaches that are extensive enough to both classify and detect SAR images. Nevertheless, with the limitation in available data for SAR images [6], employing the convolutional neural network for the SAR ATR task results in overfitting (when a model fits exactly against its training data, resulting in a poor performance against unseen data, defeating its purpose). There were three rudimentary steps taken to address this complication. The first option we call the transfer learning [7] mechanism. Here, a CNN is pre-trained using huge and extensive data before calibrating the model again for precise SAR recognition problems. However, the disparity between SAR and optical images causes low-performance accuracy in SAR Images. On the other hand, a number of unmarked SAR images

* Corresponding author.

E-mail address: qinzg@uestc.edu.cn

could be a good replacement for optical ones. The third option is image enhancement [8]. However, this method is not usually considered in most studies. Furthermore, the performance of the outcome of the studies employing image enhancements via CNNs models was not promising. This can be explained by the need to substantially strengthen the CNN Algorithm employed in these studies.

The master plan for refinement regarding the architecture of CNN generally consists of the expansion of the network's deepness and breadth. However, when the deepness is further stretched, there is a greater possibility of the network running into the challenge of vanishing/exploding gradients [9][10]. To resolve this challenge, ResNet [11] was proposed by Kaiming et al., which is composed of many residual modules that are superimposed. After the fusion of two layers, each dynamic ranges the value of the input and that of the output. Applying the principle of similarity projection makes the optimization of the variables' weighting on the network levels more rational. Additionally, this aids in stopping the problem of contours that dissipate or explode when the range is increased. For instance, in the recognition challenge utilizing the ImageNet data, the loss observed was decreased to roughly 3.57% when a deeper ResNet is made of tiers that exceed 100 [11]. Although it is known that the expansion of the depth of the network does not go on without bounds since one that is too deep is most likely to lead to overfitting. The other possible expansion is in the width of the architecture leading to the extraction of more features which is an advantage but may lead to generating more parameters and increasing the computational requirement as well as leading to overfitting.

This paper introduces a new attention-based ResNet architecture appropriate for the SAR recognition task to address this problem. This architecture focused more on extracting features because of the fewer representatives obtained from images of SAR. We summarize our key contributions as follows.

- We propose a simple channel attention mechanism for SAR ATR involving only a handful of parameters while attaining clear performance gains by eliminating discretization and using the right cross-channel interaction.
- We also explore the use of one policy learning rate in the ResNet backbone for SAR ATR.
- Finally, tests were done to see how well the proposed simple channel attention and the one policy learning rate worked on the ResNet-50 architecture for SAR ATR.

The following is how this document is organized: Section II reviews the theory of SAR ATR and attention mechanism for image recognition and classification, followed by the proposed integration of the Simple Channel Attention module in ResNet-50 architecture in Section III. Section IV provides the dataset and data preprocessing while the experimental results and analysis are seen in Section V. We concluded in Section VI.

II. RELATED WORKS

A. Introduction

Present-day major methods of classifying SAR-ATR are commonly subdivided into three methods which are template-based [12], model-based [13], and pattern-based [14]. The classic system of SAR-ATR that is template-based puts the least Mean Square Error (MSE) criteria to get the type of the target from a stored database used as a reference for the target images or templates [15]. The system that is model-based examines the detail of every image and finds out the contribution of every part of its recognition [16]. Weighed against the other two methods, the strategy that is based on the principle

of pattern recognition devoted an outstanding contribution to the task of image classification in the years past. The architecture that is pattern-based is designed for the extraction of features by initiating extractors of features which transforms the raw image to feature vectors with low dimensions. The output vectors are then categorized into groups by the classifier. A couple of ATR algorithms have seen a wide application for the classification of SAR images as well as their recognition, Artificial Neural Networks (ANN) being an example [17] with Support Vector Machine (SVM) [18] and Convolutional Neural Networks (CNN) [19] being other examples.

Not very long ago a significant surge was ignited in the field of pattern recognition by deep learning algorithms which transcended with high recognition in the interpretation of images in remoting sensing [20]. This includes recognition of SAR targets where deep learning models, such as autoencoder and CNN, have found successful applications. Knag et al. [21] used a stacked autoencoder which they developed to achieve feature fusion by applying that to SAR target classification. The utmost often used deep learning technique for SAR image classification and recognition is the CNN, with several high-content articles employing different training methods and architectures. CNN was first employed and verified by Morgan [22] for SAR Target classification. The structure of All-Convolution Networks (A-ConvNets) was proposed by this author for SAR target classification. We saw the use of CNN architecture which experimented with the MSTAR dataset for SAR target recognition in another research work [19]. The results demonstrate that the recognition rate may be considerably improved using CNN. When the convolutional layer is employed in another study [23], instead of the fully connected layer in CNN, the over-fitting concern is amazingly minimized, the parameter count is reduced, and the recognition rate is subsequently increased. Due to small samples of MSTAR datasets and overfitting, Li et al. [24] used an autoencoder to prepare the network beforehand, and the SAR images used by Jun et al. [8] were modified to enhance the sample size. Some researchers improved the network structure to improve CNN recognition performance. Zhuangzhuang [25] increased the class differentiating the performance of CNN, employed SVM for information classification, and added the class conditional independence measurement to the error cost function.

Other strategies, such as inception [26][27] and Xception [28], were put out to enhance the CNN model performance and further address the problem. The inception/X-caption techniques do not only expand the width but also split the number of channels into independent sections. The sections having varying configurations are the concatenation fusion of the feature extraction obtained from various scales so that there can be enough features acquired and work at preventing computational complexity. A network architecture that is a combination of inception module and ResNet called Inception-ResNet was proposed recently with the aim of considering both the depth and width simultaneously. Even though these techniques have been shown to improve performance for the classification of optical images, they are not applied in the field of SAR images yet. Moreover, the attributes of the images from SAR differ from those of optical images. Thus, it is theorized that it is not suitable to use methods that perform well in optical images directly for the SAR-ATR field, as such there is a need for improvements.

To further improve CNN's recognition rate and adaptability for SAR ATR, this study offers integration of simple channel attention in the ResNet-50 architecture. The simple channel attention achieves better performance by applying dimensionality reduction during learning and an appropriate cross-channel interaction to decrease model complexity. Our findings provide further evidence that our method can raise classification accuracy for the MSTAR database.

B. Attention Mechanism

A conceptual system that resembles brain activity is called the Attention Mechanism (AM) [29]. AM primarily emphasizes the important aspects while suppressing irrelevant details. With minimal cost, the AM may be added to the CNN architecture and trained alongside the CNN [30]. Attention modules vary according to their implementation ideas, such as the Convolutional Block Attention Module (CBAM) [31] which paves the path for diverse feature maps to automatically learn pixel relationships and Channel Attention Modules which create a weight matrix to assess each channel's significance. In addition to channel attention, the spatial attention module, which accumulates the weight matrix of characteristics in a spatial context, focuses on "where" relevant information might be obtained.

This study focused on the channel attention mechanism, improving deep Convolutional Neural Networks (CNNs). Nevertheless, most current approaches are intended to build more advanced modules of attention for improving performance, thereby increasing the complexity of the model. This paper proposed Simple Channel Attention (SCA) which simply requires a few arguments while attaining apparent performance gain on SAR ATR.

III. PROPOSED ARCHITECTURE

Considering that SAR image is substantially less vulnerable to reflection circumstances, overfitting is prone to happen while training CNN using SAR raw data. Since CNN is made up of huge parameters, there is severe overfitting because there aren't enough training data. By using an attention technique, this article streamlines the utilization of ResNet topology. Top-down convolutional layers gain the feature maps from the ResNet backbone network. An attention mechanism is then used to process each feature map. The results obtained using the attention mechanism are then passed through a fully connected layer that gives through the feature vectors. The final feature map is then passed through our classifier, and the classification results are acquired at the end.

A. Proposed Simple Attention Mechanism

The Channel Attention mechanism demonstrated high-performance results in improving deep CNNs. SE-Net [32] provides us with a useful method to examine channel attention and exhibits encouraging results. Therefore, the attention-module design may be classified in two ways: (1) improved feature aggregation; (2) pairing the channel and spatial attention. The proposed attention mechanism concerns the efficient convolutions designed for lightweight CNNs. Our simple channel attentions focus on the neighborhood interconnected interaction, similar to channel local attention [33] and channel-wise convolution [34]. In contrast, our approach probes a 1D convolution with adjustable Gaussian kernel size to replace fully connected layers in the channel attention module. Following the parameters of channels attention in SE Block, we assume

$$y = g(X), f_{\{w_1, w_2\}} \text{ takes the form } f_{\{w_1, w_2\}}(y) = W_2 \text{ReLU}(W_1 y) \quad (1)$$

Where $g(X) = \frac{1}{w_H} \sum_{i=1}^W \sum_{j=1}^H X_{i,j}$ denotes channel-wise global average pooling (GAP) and ReLU activation function [35]. We set the sizes of w_1 and w_2 to $c * (\frac{c}{r})$ and $(\frac{c}{r}) * c$ to prevent high model complexity. As much as Eq. (1), reducing dimensionality can minimize the model computational cost. It disrupts the weights' and the channel's straight relationship.

Both the efficiency and effectiveness of our simple channel attention mechanism can be guaranteed by using the band matrix w_k of efficient channel attention to getting the interaction of the local cross-channels. We defend the band matrix w_k thus;

$$\begin{bmatrix} w^{1,1} & \dots & w^{1,k} & 0 & 0 & \dots & \dots & 0 \\ 0 & w^{2,2} & \dots & w^{2,k+1} & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & w^{c,c-k+1} & \dots & w^{c,c} \end{bmatrix} \quad (2)$$

Where w_k in Eq. (2) involves $k * C$ parameters and the weight of y_i is computed by solely taking into account the association between k neighbors of y_i thus

$$w_i = \sigma \left(\sum_{j=1}^k w_i^j y_i^j \right), y_i^j \in \Omega_i^k \quad (3)$$

Where Ω_i^k explains k adjacent channels of y_i in sets. To distribute a constant learning rate per channel, Eq. (3) can be rewritten as follows:

$$w_i = \sigma \left(\sum_{j=1}^k w^j y_i^j \right), y_i^j \in \Omega_i^k \quad (4)$$

Which can only be executed by a fast 1D convolution with k kernel. Since our attention module is directed at capturing local cross-channel interaction, the 1D convolution kernel size k needs to be computed; thus, we adopt the below equation [5];

$$k = \psi(C) = \left\lfloor \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right\rfloor_{\text{odd}} \quad (5)$$

Where $\lfloor t \rfloor_{\text{odd}}$ denotes the nearest odd number of t . Note: we set γ and b to 3 and 1 respectively according to our experiments. Fig. 3 illustrates the implemented attention mechanism.

B. One Policy Learning Rate

The learning rate is a hyper-parameter that determines how far our network's weights are adjusted in response to the loss gradient. Conventionally, we begin training the model by gradually raising the learning rate from low to high, halting when the loss becomes uncontrollable. As a result, getting it correctly might not be easy, as shown in Fig. 1. Mathematically we have:

$$\theta_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1) \quad (6)$$

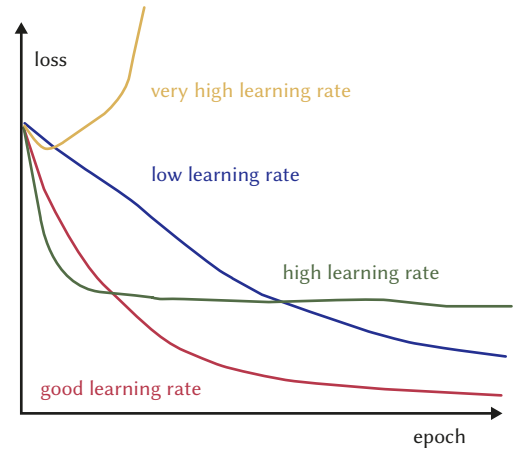


Fig. 1. Convergence effects of illustration of learning rates.

Gradient descent can be sluggish if it is too small or might overshoot the minimum if it is too great. It might either fail to converge or diverge right. Smith [36] stated that one might estimate a reasonable learning rate by first training a model with a low learning rate and then raising it (either gradually or rapidly) during every iteration, a process she called one policy learning rate. A learning rate scheduler approach enables (1) quicker network training and (2) a better understanding of the ideal learning rate. Several parameters are held constant during the experimentation, and the best learning rate is determined as the

training advances. Weight decay, maximum learning rate, optimizer, and initial learning rate are examples of such parameters, with weight decay updating the learning rate by a critical factor in each epoch.

C. Backbone Network

This paper used a lightweight deep learning network (ResNet50[11]) as its proposed model backbone, a deep convolutional neural network with a light design. It has 50 layers that, instead of learning unattributed functions, redefine as residual functions using the layer inputs. A stack of similar or “residual” blocks makes up the ResNet architecture. This block functions as a convolutional layer stack. A block’s output is also related to its input through an identity mapping mechanism. The feature mapping is continually down-sampled via depthwise convolution and the expansion in channel depth to retain the computational complexity per layer. To enable a lower

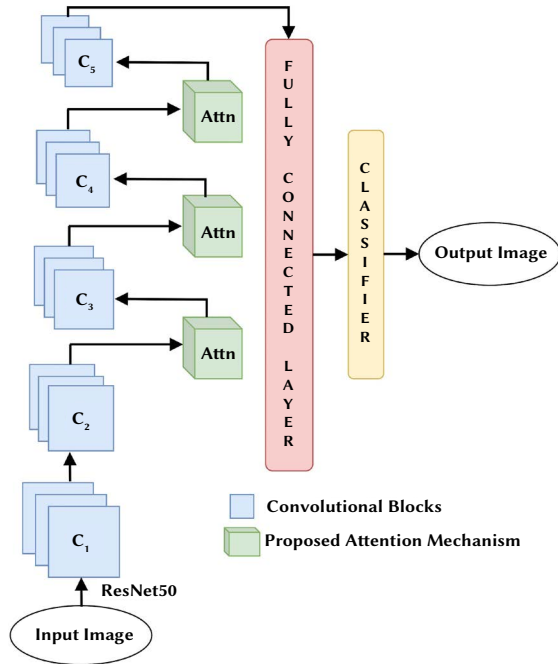


Fig. 2. Flowchart of the proposed architecture.

computational workload while computing the 3x3 convolutions in the ResNet50 model, we have a three-layer bottleneck block that employs three convolutions to reduce and restore channel depth. We denote the flow chat diagram of our proposed architecture in Fig. 2.

D. Proposed Architecture Summary

Investigating how well the squeeze-and-excitation network (SENet) performs is the suggested model’s main objective, which is the learning of channel attention to every convolution block and results in noticeable performance gains for various deep CNN architectures [37]-[40]. Although SeNet obtains higher precision, it frequently results in higher computational costs and a heavier computational complexity [11]. This paper concentrated on only three convolutional blocks while avoiding dimension reduction and accurately preserving cross-channel interaction as seen in Fig 4.

IV. EXPERIMENT

A. Dataset and Data Pre-Processing

We used the MSTAR data for our experiment and evaluations. It was created using stationary SAR and target measurements that were released by the MSTAR research and funded by the Air Force Research Laboratory (AFRL) and the Defense Advanced Research Project Agency (DARPA) [41]. It comprises ten types of tactical ground targets, as depicted in Fig. 5. The images at a 17° angle of depression were used for training while using the images at a 15° angle of depression for testing, as seen in Table I. In contrast, Table II illustrates the actual target model vs. the number of images. We used the original preprocessed data [41] in our experiment as a preprocessing technique. Before feeding to our network, all image is resized to a fixed size of 224 x 224 after some data augmentation such as random rotation and normalizing.

TABLE I. MSTAR DATASET PARTITION

	Angle	Total Number
Training Set	17°	2,752
Testing Set	15°	2,425

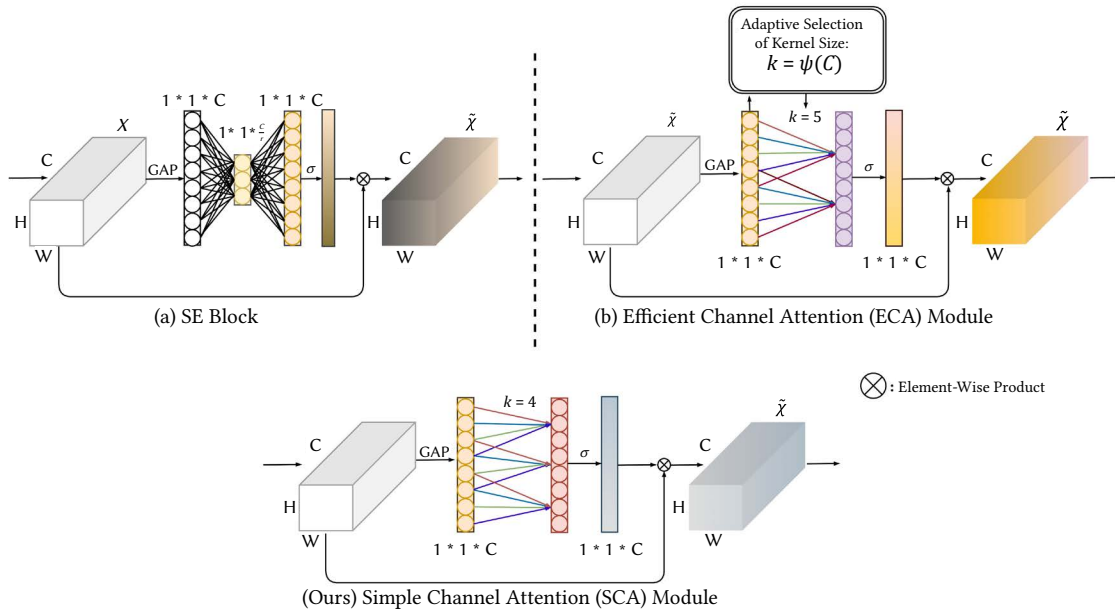


Fig. 3. An illustrative diagram of the Channel Attention Module from the SE Block to the ECA Module (The basis of our proposed SCA).

TABLE II. TARGET DESCRIPTION OF MSTAR DATABASE. THERE ARE THREE TYPES OF CLASSES THUS THE ARTILLERY CLASS, TRUCK CLASS AND THE TANK CLASS

Target Model	0	Artillery Class		Truck Class						Tank Class	
		2S1	ZSU_23_4	BRDM_2	BTR_60	SN_132	SN_9563	D7	ZIL131	T62	SN_C71
Training Set	17	300	299	299	257	233	233	300	299	299	233
Test Set	15	274	274	274	195	196	195	274	274	273	196

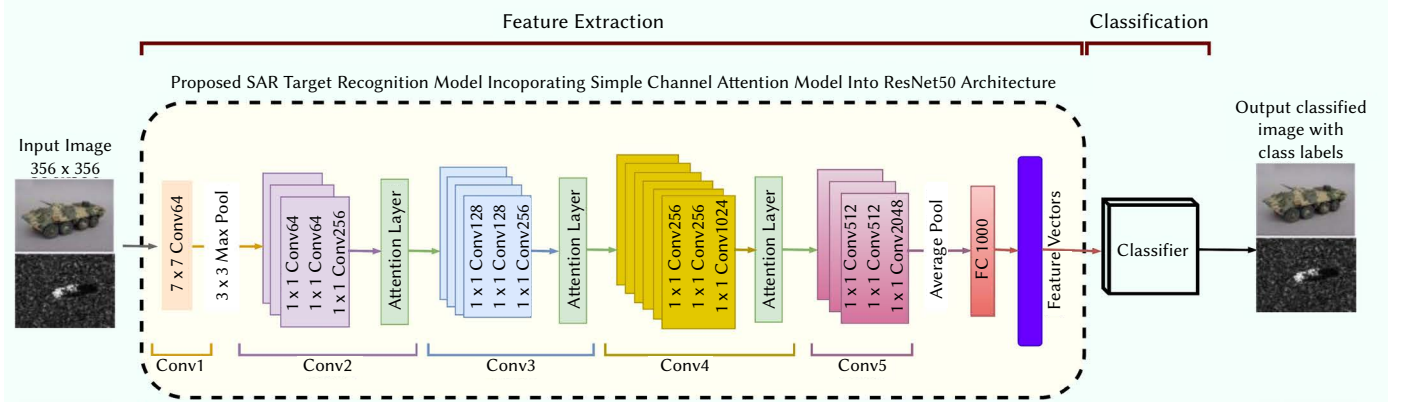


Fig. 4. Incorporating the proposed simple channel attention into the ResNet50 architecture. The proposed attention mechanism is incorporated into the second, third and fourth convolutional block to avoid the higher computational cost and computational complexity.

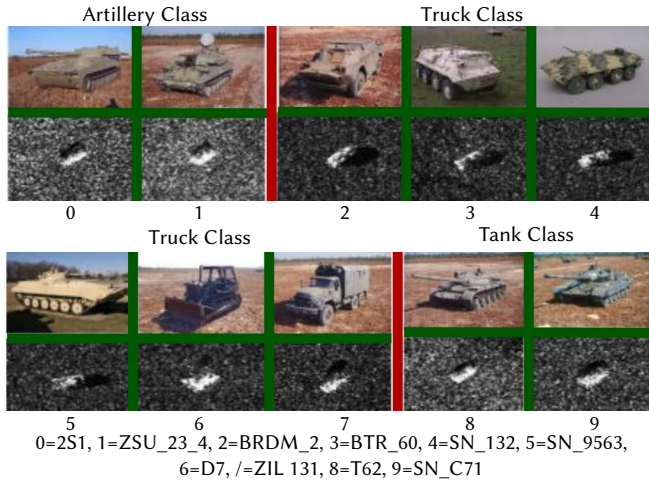


Fig. 5. Pictorial representation of the MSTAR Dataset.

B. Evaluation Metrics

General evaluation matrices including Classification Accuracy, Precision, Recall, F_1 Score, and IoU are applied in this paper. The percentage of accurately classified SAR imaging samples to all samples is used to calculate the classification accuracy. A higher percentage of correctly classified samples indicates a better classification performance. Mathematically we can express the classification accuracy as:

$$Acc = \frac{TP+FP}{TP+FP+TN+FN} \quad (7)$$

Where TP= True Positives, TN= True Negatives, FP= False Positives and FN= False Negatives.

The precision value equals ground truth SAR imagery pixels in the projected SAR imagery area divided by the number of predicted SAR Imagery pixels. The recall value is the percentage of detected SAR imagery pixels over the ground truth region. Mathematically, we express the Precision and Recall as:

$$Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN} \quad (8)$$

The F-score indicates the average overall performance as computed by precision and recall. This is how the F-Measure score is calculated mathematically:

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (9)$$

Analyzing the classification results and the loss value much further, we used the confusion matrix to envision them. It highlights the errors the classifier makes when handling multi-class situations. The predicted category is represented on the horizontal axis, while the vertical represents the correct category. Hence diagonal elements are the correctly classified SAR images. Each SAR class's classification performance is represented by its lateral elements in the standardized confusion matrix. The following illustrates how to compute the Minimum Error using the loss and variance of the ground truth and the forecasted value [42]:

$$J = -\frac{1}{N} \sum_{n=1}^N [y_n \log(\hat{y}_n) + (1 - y_n) \log(1 - \hat{y}_n)] \quad (10)$$

Where \hat{y}_n = predicted values, y_n = the ground truth, and N = number of samples. In direct contrast to accuracy, the lower the loss value, the better the model performance.

C. Implementation Details

We carried out our experiment on a windows OS computer based on the python environment, with 2.30GHz CPU Intel(R) Core (TM) i5-8300H and NVIDIA GeForce GTX 1050 Ti GPU (4g memory). We established the network using the open-source Pytorch deep learning framework, which we found to be an amazing resource. To increase our training performance, we used distributed processing relying on the CUDA 8.0 and CUDNN 5.1 prerequisites. The MSTAR dataset was used for evaluating our model. Fig. 1 displays samples of SAR images together with matching optical views. The input photos are randomly rotated horizontally and resized to 224×224 . The training hyperparameters include $1e-4$ weight decay, 0.9 momenta, 256 mini-batch, SGD optimizer, the initial learning rate of 0.1 and a reduction in learning rate of 10 per 30 epochs, 100 iterations.

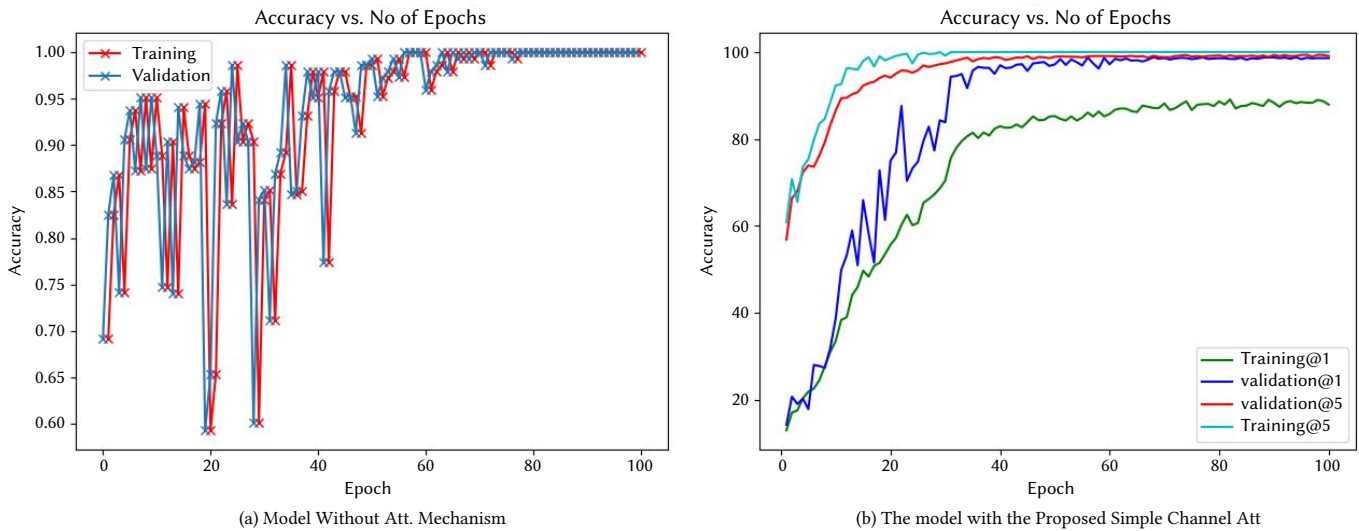


Fig. 6. Graphical representation of the training and validation Accuracy vs. Epochs. (a) shows the effects of the one policy learning rate. The models select random numbers as the learning rate until it finds a suitable range for the number for the training, which was around 60 epochs. (b) we show the training and validation accuracy at IoU 5 and 1. The model performs much better at the IoU at 5.

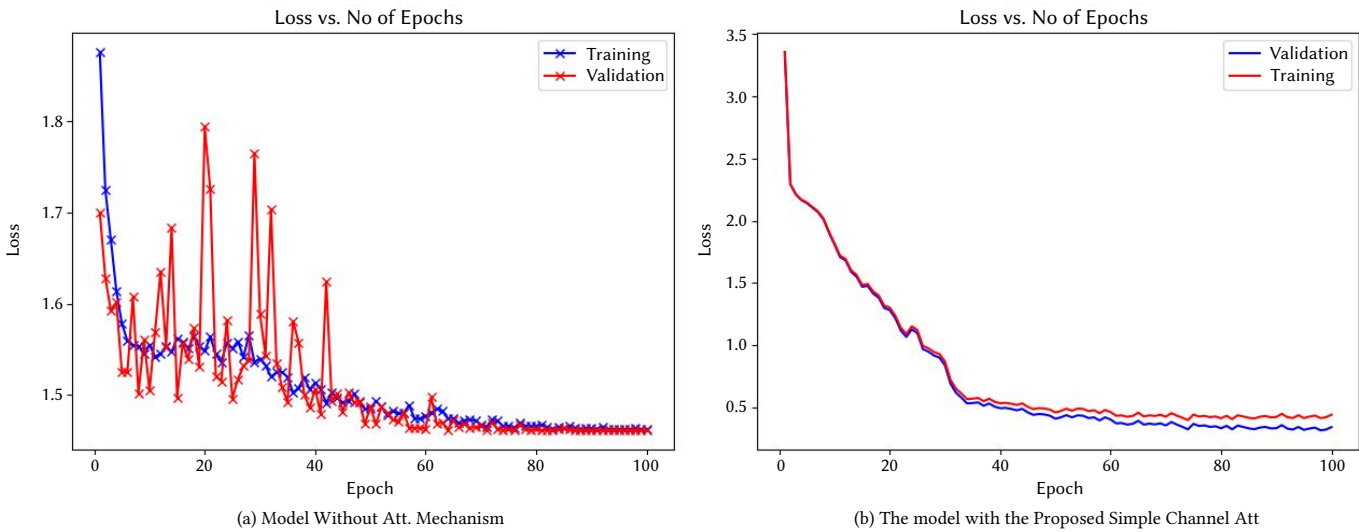


Fig. 7. Graphical representation of the training and validation Loss vs. epochs. (a) Due to the random Learning rate selection, our model losses increased until the appropriate learning rate was determined, and the loss decreased. (b) shows that the attention mechanism performs very well by choosing the correct pixels of each input for classification.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Recognition Performance Result

The proposed method is validated and discussed in this section using experimental results. First, the training and validation graph of the two proposed models is illustrated in Fig. 6, whereas Fig. 7 illustrates the training and validation loss graph. Table III. represents results from our two-model setup for the ten categories of targets. We realized that the ResNet-50 architecture based on the simple channel attention recognition rate is 100%, whereas we had a recognition rate of 99.8% in the ResNet-50 setup with one policy learning rate. As shown in Fig. 8 and Table III for the attention-based model, we obtained only 0.01 classification error in the SN₁₃₂ and SN₉₅₆₃ class under precision, SN₉₅₆₃, and SN_{C71} class under precision-Recall and finally SN₉₅₆₃ category under f1-score. Regardless of the similarities of some images in some categories, with the help of simple attention, our model could recognize the appropriate class for the test datasets.

Fig. 8 illustrates the visual performance of the proposed model against the one-policy learning rate architecture. We test using just one image from each of the MSTAR three classifications. (Artillery Class, Truck Class and the Tank Class). The first and second row depicts the simple attention mechanism and the one policy learning rate visual performance result respectively. We further undertook an empirical comparison with a few recent state-of-the-art results to validate the claims that the proposed model uses few model parameters compared to the previous work, thus attaining better results, as seen in Table IV. These CNN models have broader and deeper frameworks, and their findings are all lifted directly from the original articles. The findings above show that our proposed model outperforms benchmarked models while having substantially lower computational complexity. It is important to note that our simple attention can remarkably increase the performance of the comparable CNN models.

TABLE III. CLASSIFICATION ACCURACIES OF THE TEN CLASSES OF THE TARGET FOR THE ATTENTION-BASED MODEL VS. THE ONE POLICY LEARNING RATE-BASED MODEL

Class	Ours (Model based on an Attention Module)				Ours (Model based on One Policy Learning rate)			
	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
2S1	1.00	1.00	1.00	274	0.99	1.00	1.00	274
BRDM_2	1.00	1.00	1.00	274	1.00	0.99	0.99	274
BTR_60	1.00	1.00	1.00	195	1.00	0.94	0.97	195
D7	1.00	1.00	1.00	274	0.99	1.00	0.99	274
SN_132	0.99	1.00	1.00	196	0.97	0.99	0.98	196
SN_9563	0.99	0.99	0.99	195	0.95	0.97	0.96	195
SN_C71	1.00	0.99	1.00	196	0.98	0.99	0.98	196
T62	1.00	1.00	1.00	273	1.00	1.00	1.00	273
ZIL131	1.00	1.00	1.00	274	0.99	1.00	0.99	274
ZSU_23_4	1.00	1.00	1.00	274	1.00	1.00	1.00	274
Accuracy			1.00	2425			0.99	2425
Macro Avg	1.00	1.00	1.00	2425	0.99	0.99	0.99	2425
Weighted Avg	1.00	1.00	1.00	2425	0.99	0.99	0.99	2425

TABLE IV. MODEL PARAMETER CONTRAST BETWEEN THE PROPOSED MODEL VS. RECENT STATE-OF-THE-ART MODELS

REF	#. Param.	FLOPs	IoU@0.5	IoU@1
Ref [46]	74.45M	14.10G	-	98.18
Ref [60]	25.90M	5.36G	-	99.54
Ref [46]	46.66M	7.53G	-	98.52
Ref [42]	31.79M	5.52G	-	99.12
Ref [50]	27.35M	7.34G	-	99.18
Ref [46]	42.49M	7.35G	-	98.35
Ours	24.37M	3.86G	1.00	0.994

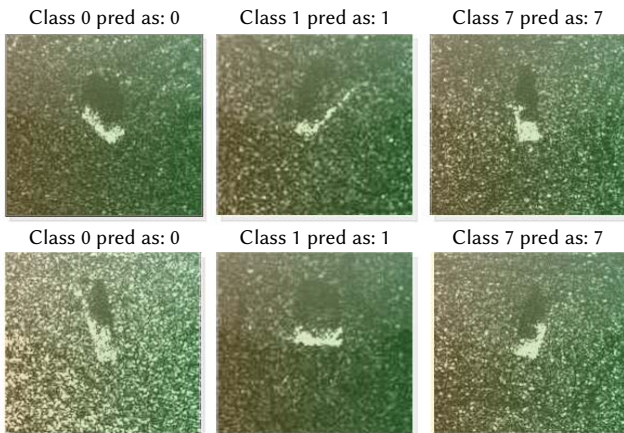


Fig. 8. Visual representation of the prediction outcome of the attention-based model vs. the one policy learning rate model.

Table III and Fig. 10 show that our one policy learning rate-based model had many misclassified samples due to similarities of the images among some classes against the attention-based model shown in Fig. 9. For the Precision, we had a misclassification rate between 0.01% - 0.05% in the 2S1, SN_132, D7, SN_9563, ZIL131 and C71 classes. For the Recall, the misclassification rate is between 0.01% - 0.06% in the BTR_60, BRDM_2, SN_9563, SN_132 and C71 classes. The F1-score had a misclassification rate between 0.01% - 0.04% in the BTR_60, BRDM_2, SN_132, D7, C71, ZIL131 and SN_9563 classes. The misclassifications result from the similarities between images in some of the classes.

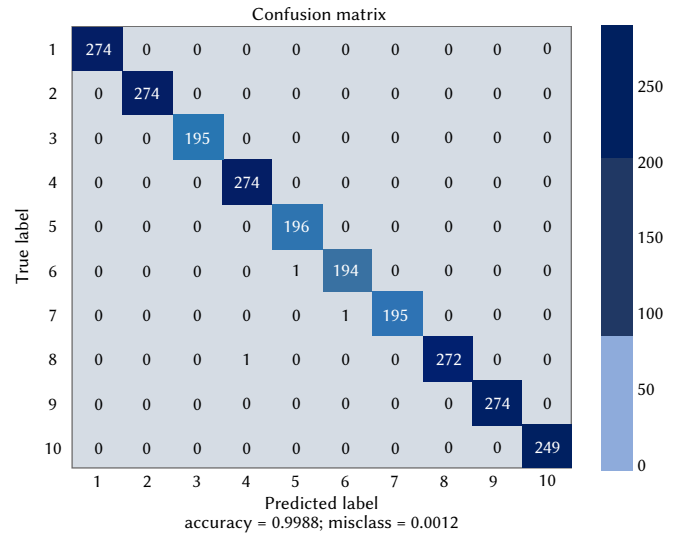


Fig. 9. Attention-Based Model Confusion Matrix of MSTAR Dataset. The accuracy of the test set is 100%.

B. Result Comparison and Discussion

As indicated in Eq. (4), our simple attention module involves a 1D convolution Kernel size denoted with K . When K is kept constant in the selected convolution blocks, our model records its best performance at $k=9$, which was obtained by an adaptive method using Eq. (5); thus, we fixed $k=9$ all through the experiment. Furthermore, the findings reveal that different deep CNNs have their best k , thus indicating that k had a positive influence on the proposed model performance. Moreover, we noted that the fluctuation of the accuracy performance between the proposed model and the one policy learning rate model was much for the one policy learning rate; thus, we concluded that deeper networks are more responsive to constant kernel size than shallower networks. Finally, substituting the SE Blocks with the primary proposed attention network with different amounts of k consistently produced superior results, demonstrating that avoiding dimension reduction and local cross-channel communication has a favorable influence on learning channel attention.

Furthermore, the proposed technique's performance is evaluated alongside 25 excellent state-of-the-art results from 2014 up to date using the same MSTAR Dataset. We pointed out the architectures used by each author in their work. We noted that our work is the

TABLE V. IOU CLASSIFICATION COMPARISON WITH THE STATE-OF-THE-ART METHODS

Author	Year	Authors Focus	IoU@0.5	IoU@1
<i>Furukawa [43]</i>	2018	End-To-End ATR of SAR Images Using Deep Learning	-	0.923
<i>Ours (One Policy Learning Rate Based)</i>	2021	Synthetic Aperture Radar Automatic Target Recognition	-	0.998
<i>Ours (Attention Based)</i>	2021	Based on Attention Mechanism	1.00	0.994

TABLE VI. STATE-OF-THE-ART RESULT COMPARISON. THE PERFORMANCE OF THE TWO IMPLEMENTED MODELS BEATS THAT OF THE STATE-OF-THE-ART MODELS. WE ANALYZED THE YEAR, THE AUTHOR'S FOCUS AND THEIR APPROACHES AND THE RESULTS OBTAINED FOR THE RECOGNITION TASK OF SAR MSTAR IMAGES

Author	Year	Authors Focus	Model	Accuracy
<i>O'Sullivan et al. [44]</i>	2001	Performance of SAR ATR with a Conditionally Gaussian Framework	Gond Gauss	97%
<i>Srinivas et al. [45]</i>	2014	Using Discriminative Graphical Models for SAR ATR	SVM	88%
<i>Dong et al. [46]</i>	2014	Using Sparse Encoding of a Single Gene Signal for ATR of SAR Images	Sparse Representation of a Monogenic Signal	93.66%
<i>Dong et al. [47]</i>	2015	Using Sparse Joint Encoding of a Single Gene Signal for ATR of SAR Images	Encoding of A Single - Gene Signal in Joint Sparse	93.41%
<i>Tian et al. [25]</i>	2016	CNN for ATR of SAR	CNN	93.76%
<i>Zhao et al. [19]</i>	2016	CNN-Based Patch Level SAR Image Classification	CNN	-
<i>Chen et al. [23]</i>	2016	Using Deep CNN for SAR Images Identification	A-ConvNet	99.13%
<i>Gorovyi et al. [48]</i>	2017	Effective SAR Images Recognition and Classification	Azimuth and Range Target Profiles Fusion	90.7%
<i>David et al. [49]</i>	2017	TL from Synthetic Data to Improve SAR ATR Models	Convnet Model	-
<i>Furukawa [43]</i>	2017	Deep Learning for SAR Image Classification Using Invariance and Data Enhancement	CNN With Data Enhancement	99.6%
<i>Chang et al. [50]</i>	2017	SAR Images ATR Based on Metadata Representations	Metadata Representation	94.88
<i>Lin et al. [6]</i>	2017	SAR Target Classification Using Deep CNN With Highway Block and Few Labeled Training Set	Deep CNN With Highway Block	99.09%
<i>Huang et al. [7]</i>	2017	TL with Deep CNN For SAR Target Recognition with Few Labeled Data	CNN-Transfer Learning	99.09%
<i>Furukawa [51]</i>	2018	End-To-End ATR of SAR Images Using Deep Learning	VersNet	99.55%
<i>Wang et al. [52]</i>	2018	CNN-Based SAR Image Target Recognition and Identification	CNN SVM	96.4% 93.85%
<i>Gao et al. [53]</i>	2018	An Improved Deep CNN Novel Algorithm for SAR Image Target Identification	DCNN + ICF + SVM	99%
<i>Dong et al. [54]</i>	2018	SAR Target Recognition Using a Salient Detail Localized Classifier Framework	Keypoint-Based Local Descriptor	-
<i>Zhang et al. [55]</i>	2019	Adaptive Region CNN for SAR Image Classification	Adaptive Neighborhood-Based CNN	-
<i>Xie et al. [56]</i>	2019	A New CNN for SAR Target Recognition	Umbrella	99.54%
<i>Xinyan et al. [57]</i>	2019	SAR Image Target Recognition with CNN	CNN	99.18%
<i>Dong et al. [58]</i>	2019	Target Recognition in SAR Images Via Dimension Reduction in The Frequency Domain	Bandwidth Modeling Approach for Sparse Signals	-
<i>Zhang et al. [55]</i>	2019	SAR Image Classification Using Adaptive Neighborhood-Based Convolutional Neural Network	Adaptive Neighborhood-Based CNN	-
<i>Wu et al. [59]</i>	2020	SAR Images ATR Based on CNN + SVM	AlexNet AlexNet + SVM Hybrid CNN Hybrid CNN +SVM	98.52% 98.35% 99.05% 99.18%
<i>Wang et al. [60]</i>	2020	SAR Target Recognition Using Recouped Non-Negative Matrix Induction and Meta-Learning	Depreciation and Amortization Non-Negative Matrix Deduction and Meta-Learning	97.9%
<i>Lie et al. [61]</i>	2021	Discrete Wavelet Transforms for Slight Discoloration in SAR Images	Contourlet-CNN	-
<i>Miao et al. [62]</i>	2021	Azimuth and Elevation Lower Bound Reconstruction for SAR Images	Adaptive Restoration with Azimuthal Sensitivity Restrictions	99.12%
<i>Ours</i>	2021	Synthetic Aperture Radar Automatic Target Recognition Based on Attention Mechanism	ResNet with Simple Attention Mechanism ResNet With One-Policy Learning Rate	100% 99.8%

first to implement an attention mechanism for the ATR SAR image recognition task; thus, we have established a new interest in research for further studies. Although the identified architectures demonstrate outstanding performance in the SAR images, as illustrated in Table V and Table VI, it is seen that the proposed architecture outperforms all the methods for SAR ATR and classification.

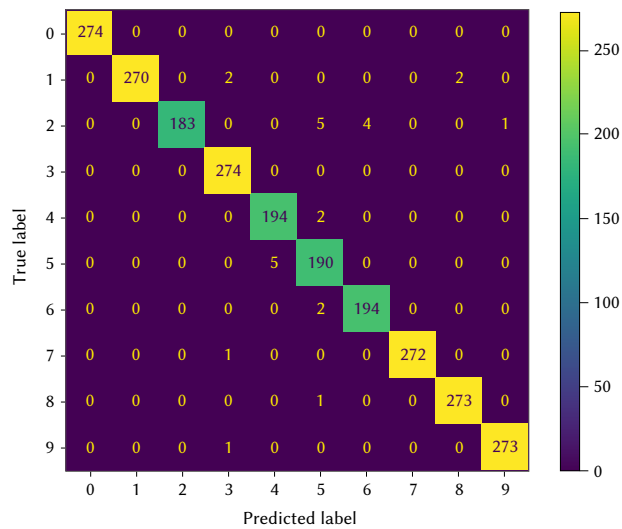


Fig. 10. One Policy Learning Rate-Based Model Confusion matrix of MSTAR Dataset. The test set yielded 99.8% accuracy.

VI. CONCLUSION

This article presents a new approach via an attention mechanism to tackle the limitation of SAR image ATR. Specifically, the channel attention mechanism is reviewed. We then proposed a simple channel attention mechanism that uses a few parameters. Yet, it yields good performance, avoids reducing dimensionality during learning, maintains cross-channel interaction performance, and decreases the complexity of the model. We fused our simple attention module into the ResNet Architecture as our network backbone. We also examined the one policy learning rate to weigh up the potential of the attention mechanism on the ResNet-50 architecture. The total identification accuracy of the ten different MSTAR SAR images is 99.8% using the one policy-based architecture and 100% using the simple attention-based architecture. Therefore, we can say that the attention-based module we created is promising to be used as a standard for SAR target identification systems.

CONFLICTS OF INTEREST

There are no conflicts of interest, according to the authors.

ACKNOWLEDGMENT

The National Natural Science Foundation of China's (NSFC) "Development of fetal heart-oriented heart sound echocardiography multimodal auxiliary diagnostic equipment" project provided funding for this study (62027827).

REFERENCES

- [1] E. M. Ampe et al., "Impact of Urban Land-Cover Classification on Groundwater Recharge Uncertainty," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 6, pp. 1859–1867, Dec. 2012, doi: 10.1109/jstars.2012.2206573.
- [2] E. P. W. Attema, G. Duchossois, and G. Kohlhammer, "ERS-1/2 SAR land applications: overview and main results," *IGARSS '98. Sensing and Managing the Environment. 1998 IEEE International Geoscience and Remote Sensing Symposium Proceedings*. (Cat. No.98CH36174), 1998, doi: 10.1109/igarss.1998.703655.
- [3] P. Gamba and M. Aldrighi, "SAR Data Classification of Urban Areas by Means of Segmentation Techniques and Ancillary Optical Data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 4, pp. 1140–1148, Aug. 2012, doi: 10.1109/jstars.2012.2195774.
- [4] D.E. Dudgeon and R.T. Lacoss. "An overview of automatic target recognition," *Te Lincoln Laboratory Journal*, vol. 6, no. 1, pp. 3–10, 1993.
- [5] Y. Cui, G. Zhou, J. Yang, and Y. Yamaguchi. "On the iterative censoring for target detection in SAR images," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 4, pp. 641–645, 2011.
- [6] Z. Lin, K. Ji, M. Kang, X. Leng, and H. Zou. "Deep convolutional highway unit network for SAR target classification with limited labeled training data," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 7, pp. 1091–1095, 2017.
- [7] Z. Huang, Z. Pan, and B. Lei. "Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data," *Remote Sensing*, vol. 9, no. 9, p. 907, 2017.
- [8] J. Ding, B. Chen, H. Liu, and M. Huang. "Convolutional neural network with data augmentation for SAR target recognition," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 3, pp. 364–368, 2016.
- [9] Y. Bengio, P. Simard, and P. Frasconi. "Learning long-term dependencies with gradient descent are difficult," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 5, no. 2, pp. 157–166, 1994.
- [10] X. Glorot and Y. Bengio. "Understanding the difficulty of training deep feedforward neural networks," *Journal of Machine Learning Research*, vol. 9, pp. 249–256, 2010.
- [11] H. Kaiming, 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 '16)*, pp. 770–778, 2015.
- [12] L.M. Kaplan "Analysis of multiplicative speckle models for template-based SAR ATR," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 37, no. 4, pp. 1424–1432, 2001.
- [13] Z. Hussein Arif et al., "Adaptive Deep Learning Detection Model for Multi-Foggy Images," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 7, pp. 26-37, 2022, doi: 10.9781/ijimai.2022.11.008.
- [14] H. Ma, J. C. Chan, T. K. Saha and C. Ekanayake. "Pattern recognition techniques and their applications for automatic classification of artificial partial discharge sources," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 20, no. 2, pp. 468–478, 2013.
- [15] G. J. Owirka, S. M. Verbout and L. M. Novak. "Template-based SAR ATR performance using different image enhancement techniques," vol. 3721 of *Proceedings of SPIE*, pp. 302–319, April 1999.
- [16] Y. Kuno, K. Ikeuchi and T. Kanade. "Model-based vision by cooperative processing of evidence and hypotheses using configuration spaces," vol. 938 of *Proceedings of SPIE*, 444 pages, Orlando, FL, USA, 1988.
- [17] S. Singha, J. T. Bellerby and O. Trieschmann. "Detection and classification of oil spill and look-alike spots from SAR imagery using an artificial neural network," in *Proceedings of the 2012 32nd IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2012*, pp. 5630–5633, Germany, July 2012
- [18] C. Yuan and D. Casasent. "MSTAR 10-Class classification and confuser and clutter rejection using SVRDM," in *Proceedings of the Defense and Security Symposium XVII*, pp. 624501–624513.
- [19] J. Zhao, W. Guo, S. Cui, Z. Zhang and W. Yu. "Convolutional neural network for SAR image classification at the patch level," in *Proceedings of the 36th IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2016*, pp. 945–948, July 2016.
- [20] X. X. Zhu, D. Tuia, L. Mou, G. S. Xia, L. Zhang, F. Xu and F. Fraundorfer. "Deep learning in remote sensing: a comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.
- [21] M. Kang, K. Ji, X. Leng, X. Xing, and H. Zou "Synthetic aperture radar target recognition with feature fusion based on a stacked autoencoder," *Sensors*, vol. 17, no. 1, p. 192, 2017.

- [22] D. A. E. Morgan. "Deep convolutional neural networks for ATR from SAR imagery," in Proceedings of the SPIE, pp. 1–13, Baltimore, MD, USA, July 2015.
- [23] S. Chen, H. Wang, F. Xu and Y. Q. Jin. "Target classification using the deep convolutional networks for SAR images," IEEE Transactions on Geoscience and Remote Sensing, vol. 47, no. 6, pp. 1685–1697, 2016.
- [24] X. Li, C. Li, P. Wang, Z. Men and H. Xu. SAR ATR based on dividing CNN into CAE and SNN. 5th Asia Pacific Conference on Synthetic Aperture Radar (APSAR), Singapore, 2015:676–679.
- [25] T. Zhuangzhuang, Z. Ronghui, H. Jiemin and Z. Jun. "SAR ATR Based on Convolutional Neural Network," Journal of Radars, vol. 5, no. 3, pp. 320–325, 2016.
- [26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich. "Going deeper with convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR' 15), pp. 1–9, IEEE, Boston, Mass, USA, June 2015.
- [27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna. "Rethinking the inception architecture for computer vision," in Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, pp. 2818–2826, July 2016.
- [28] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1800–1807, 2016.
- [29] V. Mnih, N. Heess, A. Graves and K. Kavukcuoglu. Recurrent Models of visual attention. In Proceedings of the Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 13 December 2014; pp. 2204–2212.
- [30] P. Wu, Z. Cui, Z. Gan and F. Liu. "Residual group channel and space attention network for hyperspectral image classification," Remote Sensing, vol. 12, no. 12, pp. 2035, 2020.
- [31] S. Woo, J. Park, J.Y. Lee and I. S. Kweon. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
- [32] X. Cheng, X. Li, J. Yang and Y. Tai. SESR: Single image super-resolution with a recursive squeeze and excitation networks. In 2018 24th International Conference on Pattern Recognition (ICPR) (pp. 147-152). IEEE.
- [33] D.-Q. Zhang, "cNet: Improving the Efficiency of Convolutional Neural Network Using Channel Local Convolutions," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018, doi: 10.1109/cvpr.2018.00825.
- [34] H. Gao, Z. Wang, L. Cai, and S. Ji, "ChannelNets: Compact and Efficient Convolutional Neural Networks via Channel-Wise Convolutions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 8, pp. 2570-2581, 2021, doi: 10.1109/TPAMI.2020.2975796.
- [35] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2020, doi: 10.1109/cvpr42600.2020.01155.
- [36] L. N. Smith. Cyclical Learning Rates for Training Neural Networks. Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017 pages 464 -472.
- [37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018, doi: 10.1109/cvpr.2018.00745.
- [38] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi Gather-excite: Exploiting feature context in convolutional neural networks. In NeurIPS, 2018.
- [39] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," Lecture Notes in Computer Science, pp. 3–19, 2018, doi: 10.1007/978-3-030-01234-2_1.
- [40] J. Fu et al., "Dual Attention Network for Scene Segmentation," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2019, doi: 10.1109/cvpr.2019.00326.
- [41] E. R. Keydel, S. W. Lee and J. T. Moore, "MSTAR extended operating conditions: a tutorial," in Proceedings of the Algorithms for Synthetic Aperture Radar Imagery III, vol. 2527, pp. 228–242, April 1996.
- [42] K. P. Murphy, "Machine learning: a probabilistic perspective," MIT, 2012.
- [43] H. Furukawa. Deep learning for target classification from SAR imagery: Data augmentation and translation invariance. arXiv preprint arXiv:1708.07920. 2017 Aug 26.
- [44] J. A. O'Sullivan, M.D DeVore, V. Kedia and M.I Miller, "SAR ATR performance using a conditionally Gaussian model," IEEE Transactions on Aerospace and Electronic Systems, vol. 37, no. 1, pp. 91-108, 2001.
- [45] U. Srinivas, V. Monga and G. R Raghu, "SAR automatic target recognition using discriminative graphical models," IEEE transactions on aerospace and electronic systems, vol. 50, no. 1, pp. 591-606, 2014.
- [46] G. Dong, N. Wang, and G. Kuang. "Sparse representation of monogenic signal: with application to target recognition in SAR images," IEEE Signal Processing Letters, vol. 21, no. 8, pp. 952– 956, 2014.
- [47] G. Dong, G. Kuang, N. Wang, L. Zhao and J. Lu "SAR target recognition via sparse joint representation of a monogenic signal," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 8, no. 7, pp. 3316–3328, 2015.
- [48] I. M. Gorovyi and D. S. Sharapov. Efficient object classification and recognition in SAR imagery. In 2017 18th International Radar Symposium (IRS) 2017 Jun 28 (pp. 1-7). IEEE.
- [49] D. Malmgren-Hansen, A. Kusk, J. Dall, A. A. Nielsen, R. Engholm and H. Skriver, "Improving SAR automatic target recognition models with transfer learning from simulated data," IEEE Geoscience and remote sensing Letters, vol. 14, no. 9, pp. 1484-8, 2017.
- [50] M. Chang and X. You, "Target recognition in SAR images based on information-decoupled representation," Remote Sensing, vol. 10, no. 1, pp. 138, 2018.
- [51] H. Furukawa. Deep learning for end-to-end automatic target recognition from synthetic aperture radar imagery. arXiv preprint arXiv:1801.08558. 2018 Jan 25.
- [52] Y. Wang, Y. Zhang, H. Qu and Q. Tian. Target detection and recognition based on convolutional neural network for SAR image. In 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) 2018 Oct 13 (pp. 1-5). IEEE.
- [53] F. Gao, T. Huang, J. Sun, J. Wang, A. Hussain and E. Yang, "A new SAR image target recognition algorithm based on an improved deep convolutional neural network," Cognitive Computation, vol. 11, no. 6, pp. 809-24, 2019.
- [54] G. Dong and J. Chanussot. Target Recognition in SAR Image via Keypoint-based Local Descriptor—Foundation. Preprints 2018, 2018050116 (DOI: 10.20944/preprints201805.0116.v1).
- [55] A. Zhang, X. Yang, L. Jia, J. Ai, and Z. Dong, "SAR image classification using adaptive neighborhood-based convolutional neural network," European Journal of Remote Sensing, vol. 52, no. 1, pp. 178–193, Jan. 2019, doi: 10.1080/22797254.2019.1579616.
- [56] Y. Xie, W. Dai, Z. Hu, Y. Liu, C. Li, and X. Pu, "A Novel Convolutional Neural Network Architecture for SAR Target Recognition," Journal of Sensors, vol. 2019, pp. 1–9, May 2019, doi: 10.1155/2019/1246548.
- [57] F. Xinyan and Z. Weigang, "Research on SAR Image Target Recognition Based on Convolutional Neural Network," Journal of Physics: Conference Series, vol. 1213, no. 4, p. 042019, Jun. 2019, doi: 10.1088/1742-6596/1213/4/042019.
- [58] G. Dong, H. Liu, G. Kuang, and J. Chanussot, "Target recognition in SAR images via sparse representation in the frequency domain," Pattern Recognition, vol. 96, p. 106972, Dec. 2019, doi: 10.1016/j.patcog.2019.106972.
- [59] T.-D. Wu, Y. Yen, J. H. Wang, R. J. Huang, H.-W. Lee, and H.-F. Wang, "Automatic Target Recognition in SAR Images Based on a Combination of CNN and SVM," 2020 International Workshop on Electromagnetics: Applications and Student Innovation Competition (IWEM), Aug. 2020, doi: 10.1109/iwem49354.2020.9237422.
- [60] K. Wang and G. Zhang, "SAR Target Recognition via Meta-Learning and Amortized Variational Inference," Sensors, vol. 20, no. 20, p. 5966, Oct. 2020, doi: 10.3390/s20205966.
- [61] G. Liu, H. Kang, Q. Wang, Y. Tian, and B. Wan, "Contourlet-CNN for SAR Image Despeckling," Remote Sensing, vol. 13, no. 4, p. 764, Feb. 2021, doi: 10.3390/rs13040764.
- [62] X. Miao and Y. Liu, "Target Recognition of SAR Images Based on Azimuthal Constraint Reconstruction," Scientific Programming, vol. 2021, pp. 1–10, Apr. 2021, doi: 10.1155/2021/9974723.



Professor Qin Zhiguang

Professor Qin Zhiguang is the Director of Sichuan Provincial Key Laboratory of Network and Data Security and Sichuan Next Generation Internet Data Processing Technology Engineering Laboratory. Member of the Computer Science and Technology Group of the Sixth and Seventh Academic Degrees Committee of the State Council, Member of the National Cyberspace Security First-Class Discipline Demonstration Expert Group, academic and technical leader of Sichuan Province, famous teaching teacher of Sichuan Province. Evaluation expert of the National Natural Science Foundation of China, the Ministry of Science & Technology, as well as Education. Standing Director Software Industry Association of China, director of China Cryptographic Society, chairman of Sichuan Software and Information Technology Service Industry Association, chairman of Sichuan Computer Users Association, chairman of Sichuan Software Industry Association, and vice-chairman of Sichuan Computer Society. Editorial board member of core journals such as “Computer Research and Development.”



Ukwuoma Chiagoziem Chima

In 2014, Chiagoziem C. Ukwuoma earned a B.Eng. in Mechanical Engineering majoring in Automotive Technology from the Federal University of Technology Owerri Nigeria. In 2020, he obtained his MSc. degree in Software Engineering from the University of Electronic Science and Technology of China (UESTC) where he is presently a doctoral student. His areas of interest in

research include medical imaging, attention mechanisms, and object detection and classification.



Bole Wilfried Tienin

In 2013, Mr. Bole Wilfried Tienin graduated from the Bobo-Dioulasso Polytechnic University in Burkina Faso with a Bachelor of Science in Computer Engineering. In 2018, at Turkey’s Cukurova University, he earned his MSc Degree in Computer Engineering majoring on machine learning methods for classifying, segmenting, and predicting clouds throughout his graduate studies.

At the University of Electronic Science and Technology of China, he is presently pursuing a Ph.D. (UESTC). Deep learning for target identification and recognition in Synthetic Aperture Radar (SAR) imaging is one of his main research areas.



Sophyani B. Yussif

In 2020, the University of Electronic Science and Technology of China (UESTC) awarded Mr. Sophyani B. Yussif a master’s degree in computer science and engineering majoring on Computer vision and virtual reality. At the moment, he is a doctoral degree student in Computer Science and Engineering. His previous and present free time is spent in game development in studios.



Chukwuebuka Joseph Ejayi

2014 saw Chukwuebuka Joseph Ejayi graduate with a bachelor’s degree from Nigeria’s Federal University of Technology Owerri (FUTO). In 2021, he graduated from the University of Electronic Science and Technology of China (UESTC) with a master’s degree in Software Engineering where he is presently working on his Ph.D. His research interests are in artificial intelligence, deep

learning, object recognition using a single-stage neural network, and image classification/segmentation.



Urama Gilbert Chidiebere

At the Federal University of Technology Owerri’s Department of Mechanical Engineering, Urama Gilbert Chidiebere graduated in 2014 with a B.Eng. in Mechanical Engineering—Automobile Technology. He is now pursuing his MSc in Computer Science and Technology at the University of Electronic Science and Technology of China (UESTC), with a focus on software development and machine learning. He was a member of the group that created the Agrolife Software for use by Nigerian farmers. Additionally, he developed a java-based production control system for Wsteels buildings and is now determining if blockchain technology can be used to preserve crime scene evidence.



Ukwuoma Chibueze Dabere

Ukwuoma Chibueze Dabere obtained his B.Tech. Physics Electronic from the Federal University of Technology Owerri 2018. He is currently working as a research Head and Assistant Engineer with JRB Solar investment limited company, specializing in renewable energy. He was the team leader in the Design and testing of a dual-stage rocket project in 2018—his research interest in renewable energy.



Chikwendu Ijeoma Amuche

Amuche Chikwendu Ijeoma obtained her B.Sc. in Information Management Technology from the Federal University of Technology Owerri in 2014; Masters in Information and Communication Engineering from the University of Electronic Science and Technology of China (UESTC) in 2021 where she is presently pursuing a Ph.D.

Her current work focuses on deep learning and graph representation learning, with distributed estimation and target tracking as her primary research interests.

Point Cloud Deep Learning Solution for Hand Gesture Recognition

César Osimani¹, Juan Jesus Ojeda-Castelo², Jose A. Piedra-Fernandez² *

¹ Applied Research & Development Center on IT (CIADE-IT) Universidad Blas Pascal, Córdoba (Argentina)

² Applied Computing Group (ACG), Department of Informatics University of Almeria, Almeria (Spain)

Received 11 March 2021 | Accepted 11 March 2022 | Published 10 January 2023



ABSTRACT

In the last couple of years, there has been an increasing need for Human-Computer Interaction (HCI) systems that do not require touching the devices to control them, such as ATMs, self service kiosks in airports, terminals in public offices, among others. The use of hand gestures offers a natural alternative to achieve control without touching the devices. This paper presents a solution that allows the recognition of hand gestures by analyzing three-dimensional landmarks using deep learning. These landmarks are extracted by using a model created with machine learning techniques from a single standard RGB camera in order to define the skeleton of the hand with 21 landmarks distributed as follows: one on the wrist and four on each finger. This study proposes a deep neural network that was trained with 9 gestures receiving as input the 21 points of the hand. One of the main contributions, that considerably improves the performance, is a first layer of normalization and transformation of the landmarks. In our experimental analysis, we reach an accuracy of 99.87% recognizing of 9 hand gestures.

KEYWORDS

Artificial Neural Network, Computer Vision, Hand Gesture Recognition, Point Cloud.

DOI: 10.9781/ijimai.2023.01.001

I. INTRODUCTION

THERE is a high interest in using LiDAR scanners (Light Detection and Ranging) which use beams of light to measure distance to objects, allowing to acquire a three-dimensional point cloud of the environment [1]. The information acquired by this type of scanner combined with object color information is interesting for several applications (e.g., construction of three-dimensional models from the scanning of real objects [2], identification of objects within an environment [3], or self-driving cars [4]). Devices that combine color information (standard RGB cameras) and the data obtained by LiDAR scanners are more often called depth cameras or D-RGB (Depth - Red Green Blue) sensors. Among them we can find Kinect for Windows, Leap Motion Controller o Intel RealSense, which can be found in offices and homes as they are affordable. However, they are not consumer devices, as RGB cameras are.

If we get into the topic of Human-Computer Interaction and the constant effort to incorporate increasingly natural interactions, we find the commands by voice or through gestures of the face, body or hands. Let's focus on Computer Vision and the area of study related to hand gestures, particularly to one aimed at controlling Natural User Interfaces (NUI).

The identification of hand gestures can be interesting to create user interfaces with the aim of achieving better experiences, such as in augmented reality applications [5] overlapping virtual contents or digital information in an aligned way with the real image of the hand or applications to control devices. This identification of hand gestures is not trivial considering the hands and their fingers are, generally, occluded from each other, and their contours do not have high contrast.

In this work we propose the identification of 9 hand gestures by interpreting a cloud of 3D reference points obtained through a standard RGB camera. We introduce a neural network architecture which has the follow main advantages: a small number of hidden layers and high prediction hit rate of hand gestures. In this way, we achieve good results in predictions and the possibility of working on CPU not only to make predictions but also to train the network.

The rest of the paper is organized as follows: section II describes the Related Work, section III explains our Proposed Work, section IV explains the results and section V includes the conclusions.

A. Contributions

A deep learning model has been developed to recognize 9 hand gestures by analyzing a point cloud of sparse 3D landmarks of the hand. The network architecture has at its input a transformation and normalization layer that allows achieving very good classification hit rates, even when using third party datasets containing different user profiles and variable environments.

* Corresponding author.

E-mail addresses: cosimani@ubp.edu.ar (C. Osimani), juanje.ojeda@ual.es (J. J. Ojeda-Castelo), jpiedra@ual.es (J. A. Piedra-Fernandez).

II. RELATED WORK

A. Point Cloud

A point cloud is a fancy name for a group of points in space (here we will refer to three-dimensional space, but the concept is extensible to any dimension). There are different ways to collect point clouds from the objects that exist in an environment, among the most common are LiDAR scanners, depth cameras or some models created with automatic learning to infer reference points for hands [6], faces [7] or skeletons of bodies [8].

Point clouds have been applied mainly in detecting objects as shown in the works described below. In [9] a framework called PointRCNN has been developed to detect 3D objects through point clouds. This framework consists of 2 phases: in the first one 3D bounding boxes are used to generate segmentation masks in a bottom-up architecture. The second phase is essential to improve the efficiency of this approach with the combination of semantic and local spatial features. In [10] VoxelNet is presented, which is a deep network to perform 3D detections, with the particularity of joining the feature extraction processes and the prediction of 3D bounding boxes in one phase, unlike PointRCNN where they were carried out in 2 phases. One of the main advantages of VoxelNet is that it does not perform hand-crafted feature extraction, which can be understood as features extracted from separate images according to a certain manually predefined algorithm based on the knowledge of experts. Features extracted with Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG) are commonly known examples of hand-crafted features. Although the previous cases allowed to perform object recognition in a generic way, studies have also been done to focus on the detection of a specific object, that is the case of this work [11] where point cloud data has been applied to perform a vehicle detection in order to integrate it into an autonomous driving system. To achieve this goal, the authors have proposed a 3D convolutional network to improve performance in the point cloud detection task. However, they have also been used in gesture recognition, i.e., in [12] a recognition of hand gestures based on 3D and 2D representations to control a virtual world in 3D was proposed. The 3D features are based on the finger position in the point cloud, while the 2D features come from the outline of the hand drawn from a series of images. This system has the outstanding characteristics that it can recognize both static and dynamic gestures, where the algorithm used to classify static gestures has been Support Vector Machine, while Dynamic Time Warping has been used for dynamic gestures. In addition, in the evaluation process of this work, a 95% success rate was obtained for static gestures and 81.34% for dynamic gestures.

B. Classification and Segmentation of Point Cloud

Once a point cloud has been collected, it may be necessary to isolate the different objects, that is, to perform a segmentation, or also to classify each of those objects. Deep Learning has a good performance for classification of point clouds and this is demonstrated by the Multilayer Perceptron (MLP) called PointNet [13] that achieves an accuracy of between 80% and 90% for classification of point clouds using the dataset ModelNet40 [14] which contains 40 classes of objects such as chairs, desks, beds, tables and others. The point cloud of a chair is shown in Fig. 1.

PointNet has been applied in many studies, some examples are described then. This work [15] aims to improve PointNet to increase object classification performance which is the main use of this model. To achieve this objective, two actions have been carried out: one is to increase the number of hidden layers of the architecture and the other is to combine the softmax loss function with center loss. In this way, an accuracy of 89.95% has been obtained. In [16] the PointNet



Fig. 1. Point cloud of a chair.

network has been trained in order to verify the performance of this deep network in the human body segmentation task. To perform the segmentation in PointNet, the SMPL model is used, which offers a realistic 3D model of the human body. In this work two types of tasks have been approached: a segmentation and a classification task. In each task a different simplification of the PointNet architecture has been used. In the segmentation task, the points that have been located on the surface of the body are obtained, while in the classification task a binary classification is carried out to identify the body of a man and a woman. On the assumption of gesture recognition, Ge et al [17] propose a PointNet that has the purpose of processing the 3D point clouds to obtain a representation of the pose of the hand in 3D. This system is based on analyzing the 3D point cloud to obtain an estimate of the joints of the hand in 3D and to get better results, the points have been normalized so that it is insensitive to the variations that may arise from the location of each one of them. Furthermore, it has been possible to improve the precision of the position of the fingertips using a PointNet that obtains the neighboring points of the estimation of the location of the fingertips, having as a consequence that the model is more robust.

Among jobs that address the challenge of unsupervised learning with point clouds, we can find FoldingNet [18] and PointCNN [19].

In the same line appears PointNet++ [20] that adds a neighborhood of points to capture features that allow to group close points.

There are also works, such as [21], that improve the segmentation of the different objects in a point cloud by processing point clouds within a temporary space, that is, point clouds obtained from multiple instants of consecutive times.

Regarding the work that adjusts, aligns and superimposes a 3D model of a hand on the hand detected in a single image, we can find [22], which also proposes an approach to the automated collection of data from Youtube to incorporate them into the dataset in order to include data unrelated to the laboratory.

C. Deep Learning in Gesture Recognition

PointNet is a deep network that has been used in this work to perform gesture recognition, but there are other proposals in Deep Learning that have also been applied to perform this type of recognition. In [23] the aim was to develop a framework to recognize human actions applying the Convolutional Neural Network (CNN). This system consists of two phases. In the first one the activities that involve single-limb are separated from those that are multi-limb to perform the classification of said activities in the next phase. In the classification stage, two CNNs were used to detect the two types of activities that were identified in the previous phase, obtaining a 97.88% hit rate. Khari et al [24] use learning transfer to do static gesture recognition. In this study, the VGG19 model has been trained with RGB and RGB-D images to identify of 24 gestures from the ASL dataset. This proposal has been compared with other models such as VGG16, CaffeNet or Inception V3, being the presented proposal in this work the one with the highest hit rate with 94.8%.

Another type of gesture recognition is based on using devices or sensors, which provide a set of data that are useful for such recognition [25]. deepGesture [26] is a methodology that recognizes gestures with the arm through the data it receives from the gyroscope and accelerometer of an arm band using Convolutional Neural Network and Recurrent Neural Network. In this process, the input data obtained from the arm band are entered in the Convolutional Neural Network to extract the characteristics and then in the layers of the Recurrent Neural Network to improve the performance of gesture recognition, which has improved the precision of each class by 6%.

III. PROPOSED WORK

The aim of our work is to achieve a hand gesture detection model that allows developing solutions to control devices (such as a graphical user interface, virtual keyboard or mouse) in a natural and intuitive way. In the following we will detail the steps we follow in order to approach the solution. The steps we will detail below are the following: obtaining a point cloud of the hand, choosing the gestures to be used, creating the dataset, normalizing the data, defining the network architecture, training and obtaining the model for the predictions.

A. Inference of KeyPoints

This work implements the model *Mediapipe hands* [6] created with automatic learning techniques to infer 21 three-dimensional reference points of a hand from the processing of a single image. These 21 points (from now on KeyPoints) are located: one on the wrist and 4 more points on each finger (as shown in Fig. 2).

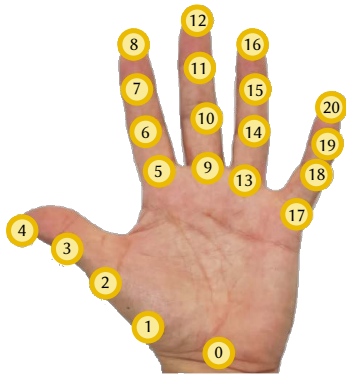


Fig. 2. KeyPoints of a hand.

B. Gesture Selection

In order to make a selection of gestures that users can choose from, we have explored gestures from non-verbal communication, sign language and related articles to Human-Computer Interaction. A variety of hand gestures are used in natural user interfaces and it is common to find solutions that use: the tip of the index finger or the open palm of the hand to control the mouse; the closed hand (fist) followed by the open hand to drag & drop; the thumb up to accept or the thumb down to cancel.

We want to obtain a model that recognizes a set of gestures to be able to design solutions in the future where users can select one by one the gestures for different actions (such as clicking the mouse, scrolling, moving the mouse pointer, moving forward or backward in a presentation, accepting or canceling). Just by obtaining an identifier for each gesture, either a letter or a number, we explored different sign languages and selected the following (visualized in the Fig. 3):

- From International Sign language: 1, 4, 5
- From American Sign Language: 9, V, W
- From French Sign Language: A, L, S



Fig. 3. Gesture names: S - 1 - V - W - 4 - 5 - 9 - L - A.

C. Dataset

In order to create our dataset, we requested video recordings from 10 volunteers. Each of them recorded a single video of approximately 3 minutes performing the 9 gestures without interrupting the recording. All movements were executed under free style, speed and direction to the personal liking. Subsequently, all the videos were processed in order to extract a sequence of grouped and annotated images for each gesture. A total of 39,150 images were obtained in a balanced way between gestures and volunteers. The *Mediapipe Hands* [6] model was used to extract the 21 keypoints of the hands from the complete set of images. A couple of sample images of this dataset with its detected KeyPoints are shown in Fig. 4.



Fig. 4. Example of images to extract KeyPoints.

The dataset consists of a CSV (comma-separated values) file containing 39,150 records (4,350 for each of the 9 gestures) with the information shown in Fig. 5. Each record has 64 columns of information: the name of the gesture plus 21 KeyPoints (x, y, z).

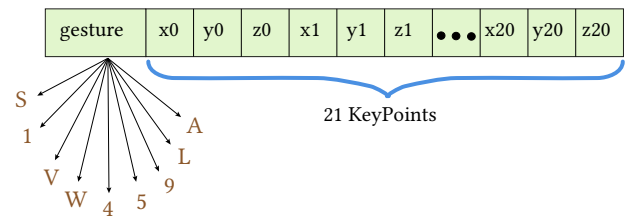


Fig. 5. Stored information.

This dataset is divided into a proportion of 80% for training and validation data (31,320 samples), and 20% for testing (7,830 samples).

In addition, we have downloaded 3 external datasets [27], [28] and [29] to test our model. Since these datasets do not contain our 9 gestures, we have combined them to reach a set of 4,500 samples (500 for each gesture).

D. Data Normalization

After generating the dataset of 39,150 images, data normalization is performed, which consists of several transformations (translation, rotation and scaling) so that KeyPoints are located at the origin of three-dimensional space and the middle finger aligned with Y-axis.

Following transformation matrices are used for normalization:

- Matrix (1) to translate to origin.

$$T = \begin{bmatrix} 1 & 0 & 0 & -kp0x \\ 0 & 1 & 0 & -kp0y \\ 0 & 0 & 1 & -kp0z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1}$$

where KeyPoint 0 is $(x, y, z) = (kp0x, kp0y, kp0z)$

- Rotation matrix (2) around an arbitrary axis: To align the middle finger with Y-axis.

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} & 0 \\ r_{21} & r_{22} & r_{23} & 0 \\ r_{31} & r_{32} & r_{33} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2)$$

where:

$$\begin{aligned} r_{11} &= \cos \theta + u_x^2(1 - \cos \theta) \\ r_{12} &= u_x u_y(1 - \cos \theta) - u_z \sin \theta \\ r_{13} &= u_x u_z(1 - \cos \theta) + u_y \sin \theta \\ r_{21} &= u_y u_z(1 - \cos \theta) + u_x \sin \theta \\ r_{22} &= \cos \theta + u_y^2(1 - \cos \theta) \\ r_{23} &= u_y u_z(1 - \cos \theta) - u_x \sin \theta \\ r_{31} &= u_z u_x(1 - \cos \theta) - u_y \sin \theta \\ r_{32} &= u_z u_y(1 - \cos \theta) + u_x \sin \theta \\ r_{33} &= \cos \theta + u_z^2(1 - \cos \theta) \\ r_{33} &= \cos \theta + u_z^2(1 - \cos \theta) \end{aligned}$$

Here, u is a unit vector that is perpendicular to the plane formed by KeyPoint 9 vector and Y-axis. Knowing that two vectors are perpendicular (or orthogonal) when their dot product (or scalar product) is equal to zero, then we can calculate the vector u . Or it is also possible to use the vector product (or cross product) between KeyPoint 9 and Y-axis. To do this, we can use the Rule of Sarrus to calculate the 3×3 determinant and thus obtain a vector perpendicular to the plane between KeyPoint 9 and Y-axis. Finally, we divide it by the norm to obtain a unit vector which is the vector u of the previous expressions.

By Rule of Sarrus we obtain a vector (Eq. 3) that, in general, is not unitary. We consider that the vector on the Y-axis is unitary, that is, it is the vector $(0, 1, 0)$:

$$\begin{aligned} u'_x &= +(kp9y * 0 - 1 * kp9z) = -kp9z \\ u'_y &= -(kp9x * 0 - 0 * kp9z) = 0 \\ u'_z &= +(kp9x * 1 - 0 * kp9y) = +kp9x \end{aligned} \quad (3)$$

When we divide by its norm we get the unit vector u , as shown in Eq. 4.

$$\begin{aligned} |u'| &= \sqrt{(-kp9z)^2 + 0 + (kp9x)^2} \\ u &= \left(\frac{-kp9z}{|u'|}, 0, \frac{kp9x}{|u'|} \right) \end{aligned} \quad (4)$$

θ is the angle between the vector formed from the origin to the start of the middle finger (i.e. KeyPoint 9) and the unit vector on the Y-axis. It can be calculated as given in Eq. 5.

$$\theta = \cos^{-1} \left(\frac{kp9y}{\sqrt{(kp9x)^2 + (kp9y)^2 + (kp9z)^2}} \right) \quad (5)$$

- Matrix (6) to rotate palm on the Y-axis so it is aligned with plane $z = 0$.

$$R_y = \begin{bmatrix} \cos \beta & 0 & \sin \beta & 0 \\ 0 & 1 & 0 & 0 \\ -\sin \beta & 0 & \cos \beta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (6)$$

β is the angle on the plane $y = 0$ of the angle formed between KeyPoint 17 and X-axis. In this way, we align the palm with the plane $z = 0$ as shown in Eq. 7.

$$\beta = \tan^{-1} \left(\frac{kp17z}{kp17x} \right) \quad (7)$$

- Rotation matrix on Y-axis to place the palm in a frontal way: we use the R_y matrix to rotate 180° over Y-axis as long as the palm is

in the direction of the negative values of z . To know if the palm is facing forward or not, a simple calculation is done by detecting where fingertips are facing.

- Mirror with respect to the plane $x = 0$: Regardless of whether it is right or left hand, we want to mirror the hand in such a way that the thumb always remains towards positive values of x . It is easy to detect if the thumb is to the right or to the left by finding out x values of the KeyPoints belonging to the thumb.
- Scaling: The hand is scaled in order that the magnitude $|kp0y - kp9y|$ is equal to 100. The matrix (8) is used to solve it.

$$E = \begin{bmatrix} \frac{100}{kp9y} & 0 & 0 & 0 \\ 0 & \frac{100}{kp9y} & 0 & 0 \\ 0 & 0 & \frac{100}{kp9y} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (8)$$

To obtain normalized values, operations (shown in Eq. 9) are performed with these matrices with each of the 21 KeyPoints of each hand.

$$\begin{bmatrix} x_n \\ y_n \\ z_n \\ 1 \end{bmatrix} = ER_y RT \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (9)$$

where x_n , y_n and z_n are the coordinates of the normalized KeyPoints. Bear in mind that, depending on the case, the 180° rotation and/or the mirror with respect to the plane $x = 0$ is also carried out.

To carry out normalization of the KeyPoints, we developed a tool that allows visualizing the correct normalization of KeyPoints that make up our dataset. In Fig. 6, a hand is shown in its original position and in Fig. 7 it is displayed after normalization. Some KeyPoints were joined with lines for a clear visualization of the hand's skeleton.

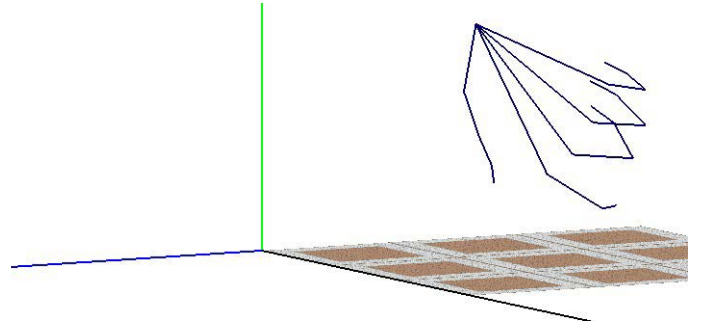


Fig. 6. Skeleton of a hand in its original position.

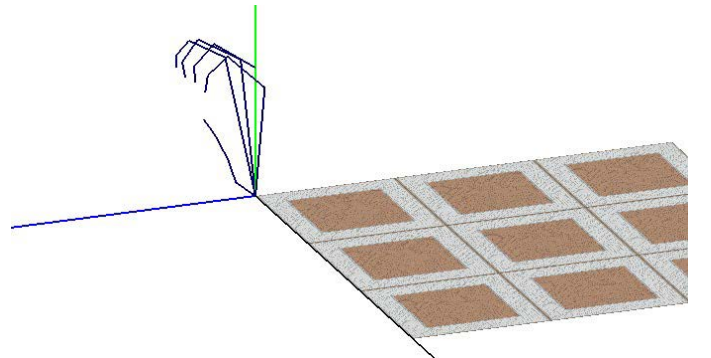


Fig. 7. Skeleton of a hand after normalization.

E. PointNet Network Architecture

There is a type of neural network called PointNet [30], which receives in its input layer a point cloud for object classification. In general, point clouds are obtained from objects in an environment, and the challenge is to be able to classify and/or segment each one of them from this point cloud, which is just a bunch of isolated points that vaguely describe the structures and surfaces of the objects. The following is a brief discussion of some characteristics when classifying a point cloud:

- Invariance to permutation: a point cloud is a set of raw data, without additional information. It is a collection of (x, y, z) coordinates without structure. This makes the data invariant to permutations.
- Invariance to transformations: the classification of objects should not change if the point cloud undergoes translation and/or rotation transformations (not so with scaling).
- Importance between neighboring points: each point is not treated independently as the interaction between neighboring points contains useful information.

It is important to note that it is common to consider that a point cloud has a large number of points. The PointNet authors used in their work a cloud of 2048 points for each object, using the ModelNet10 dataset [14], which contains objects belonging to 10 classes.

PointNet network architecture for classification of a point cloud can be visualized and analyzed in [30]. This network takes n entry points, each one with dimension 3 belonging to (x, y, z) coordinates. The authors propose 2048 points for each object, so it would have an input with dimension [2048, 3]. It has two groups of layer called T-Net which are also neural networks that perform transformations on the data without modifying its dimension. These T-Net subnets are composed of temporal convolutions (Conv1D) with ReLU activation, batch normalization, 1D Max Pooling and densely connected layers (Fully Connected).

After transformations with T-Net combined with the convolution layers, a Max Pooling (GlobalMaxPooling1D) is performed, taking the global maximum value of the data, decreasing the dimensionality. It is followed by Fully Connected Layers, Dropout layers and a last layer with softmax activation function to obtain the scores for k output classes. PointNet network uses optimization with Adam stochastic gradient descent method and cross entropy as loss function. We analyze this network architecture and propose some modifications which are discussed below.

F. Modified Network Architecture

T-Net subnets perform affine transformations in data and we propose to eliminate them, since our dataset already has different transformations that apply a normalization. We also propose to include new convolution layers and Fully Connected Layers, leaving an architecture as shown in Fig. 8, which was one of the best results. Note that the data normalization explained above is carried out beforehand.

G. Data Increment

While analyzing a graph of 21 KeyPoints of a hand it can be difficult, to the human eye, to identify to which gesture those points correspond. It can be considered that 21 KeyPoints are insufficient to represent the skeleton of a hand, so we can generate extra data by knowing that among certain KeyPoints there is a hand bone (in the palm the metacarpal bones, in the beginning of the fingers the proximal phalanges, followed by the middle phalanges and at the tip of the fingers the distal phalanges).

Layer	Output Shape	Param #
InputLayer	(None, 21, 3)	0
Conv1D	(None, 21, 32)	128
BatchNormalization	(None, 21, 32)	128
Activation	(None, 21, 32)	0
Conv1D	(None, 21, 64)	2112
BatchNormalization	(None, 21, 64)	256
Activation	(None, 21, 64)	0
GlobalMaxPooling1D	(None, 64)	0
Dense	(None, 128)	8320
BatchNormalization	(None, 128)	512
Activation	(None, 128)	0
Dropout	(None, 128)	0
Dense	(None, 7)	903
Total params: 12,359		
Trainable params: 11,911		
Non-trainable params: 448		

Fig. 8. Proposed network architecture.

In this regard, a new parameter is defined which allows to incorporate a certain amount of additional KeyPoints on the bones of the hand. This is achieved by calculating lines that join the KeyPoints that correspond to the ends of the bones mentioned above. In Fig. 9 we can see KeyPoints of a hand with the addition of 10 KeyPoints on each bone.

H. Training

At this point we have the network architecture defined with the Keras library and the dataset with 31,320 samples for training and validation. We continue with the training in order to obtain a model (in HDF5 format) that allows us to make predictions.



Fig. 9. Hand with 10 extra KeyPoints on each bone.

Keep in mind that we have a total of 39,150 samples in our own dataset plus 4,500 samples obtained from third-party image datasets. From the 39,150 samples, we separated 31,320 for training and validation, and 7,830 for testing. Note that we have two sets of samples for testing, one of which was randomly sampled from our own dataset and the other has been generated from third-party images.

IV. EXPERIMENTS AND EVALUATION















A. Performance of the Proposed Network

Several network trainings were performed modifying parameters such as the number of epochs, learning rate of the Adam optimization method and the number of extra KeyPoints on each bone. In Table I are

TABLE I. 7,830 PREDICTIONS MADE

#	Extra KeyPoints on each bone	Learning rate	Epochs	Total KeyPoints	training loss	training acc	validation loss	validation acc	Correct predictions (ext dataset)	Correct predictions (%) (ext dataset)	Correct predictions (own dataset)	Correct predictions (%) (own dataset)
1	0	0.0005	10	21	0.0068	99.84 %	0.0048	99.95 %	4336 of 4500	96.36 %	7820 of 7830	99.87 %
2	2	0.0005	30	61	0.0085	99.80 %	0.0056	99.89 %	4305 of 4500	95.67 %	7820 of 7830	99.87 %
3	10	0.001	50	221	0.0062	99.85 %	0.0094	99.89 %	4297 of 4500	95.49 %	7820 of 7830	99.87 %
4	0	0.0005	30	21	0.0055	99.88 %	0.0072	99.86 %	4305 of 4500	95.67 %	7819 of 7830	99.86 %
5	5	0.0005	30	121	0.0054	99.90 %	0.0038	99.92 %	4349 of 4500	96.64 %	7819 of 7830	99.86 %
6	0	0.001	30	21	0.0070	99.81 %	0.0027	99.92 %	4315 of 4500	95.89 %	7819 of 7830	99.86 %
7	5	0.001	20	121	0.0084	99.80 %	0.0035	99.92 %	4340 of 4500	96.44 %	7819 of 7830	99.86 %
8	5	0.001	30	121	0.0064	99.86 %	0.0039	99.94 %	4289 of 4500	95.31 %	7819 of 7830	99.86 %
9	10	0.001	30	221	0.0075	99.82 %	0.0073	99.90 %	4296 of 4500	95.47 %	7819 of 7830	99.86 %
10	0	0.0005	50	21	0.0058	99.85 %	0.0050	99.89 %	4320 of 4500	96.00 %	7818 of 7830	99.85 %

TABLE II. NUMBER OF WRONG PREDICTIONS BY MODEL 1

Said...	?	It was...	Number of wrong predictions
 L		1	1
 L		A	1
 5		9	1
 V		W	1
 4		W	2
 1		V	2
 S		A	2

shown the results ordered according to successes in predictions made with the test dataset of KeyPoints belonging to 7,830 hand samples and, in addition, 4500 samples of external datasets. The table shows the following: the amount of additional keyPoints added on each bone; the learning rate of Adam optimizer; the amount of epochs for training with batch size of 32 with a division of 80%/20% for training/validation; the total of KeyPoints for each hand; the loss in the training set after all the epochs; the accuracy with the training set; the loss in validation set after all epochs; the accuracy in validation set; the success rate in predictions made with a test set of 4,500 samples of external datasets; and the success rate in predictions made with a test set of 7,830 own samples (independent of the train/val set). The time consumed to perform each prediction is 26 mil-liseconds on average on an Intel® Core™ i7-1165G7 Processor (without GPU).

B. Metrics

In order to observe the performance of the proposed architecture we will resort to analysis of trained model number 1, 5 and 10 presented in Table I.

- **Model number 1:** In order to have a quick approximation to the performance of this model, let's analyze the confusion matrix in

TABLE III. METRICS FOR TRAINING NUMBER 1

gesture	precision	recall	f1-score	support
s	0.9977	1.0000	0.9989	870
1	0.9977	0.9989	0.9983	870
v	0.9988	0.9977	0.9983	870
w	1.0000	0.9966	0.9983	870
4	0.9977	1.0000	0.9989	870
5	0.9989	1.0000	0.9994	870
9	1.0000	0.9989	0.9994	870
L	0.9977	1.0000	0.9989	870
a	1.0000	0.9966	0.9983	870

accuracy = 0.9987 for 7830 predictions (870 each class)

Fig. 10. Each column represents the number of predictions made by this model for each of the 9 gestures, while rows represent the true gesture. For these predictions, the own test set composed of 7,830 samples (870 for each gesture) is used, which is independent of set used for training and validation.

s	870	0	0	0	0	0	0	0	0
1	0	869	0	0	0	0	0	1	0
v	0	2	868	0	0	0	0	0	0
w	0	0	1	867	2	0	0	0	0
4	0	0	0	0	870	0	0	0	0
5	0	0	0	0	0	870	0	0	0
9	0	0	0	0	0	1	869	0	0
L	0	0	0	0	0	0	0	870	0
a	2	0	0	0	0	0	0	1	867
	s	1	v	w	4	5	9	L	a

Fig. 10. Confusion matrix of model number 1.

By breaking down a little the information of the confusion matrix, we can observe the incorrect predictions in Table II where it is shown in the first column the gesture predicted by the model,

in the second column the true gesture and in the third column the number of times that there was confusion. In Table III are presented metrics that mean the following:

- **Precision:** It provides information about false positives, as shown in Eq. 10. It is the ratio between well classified positive cases and the total number of predictions made.

$$\text{precision} = \frac{TP}{TP + FP} \quad (10)$$

where:

TP is the number of true positives

FP is the number of false positives.

- **Recall:** It indicates the ratio of positive classes that the model has been able to predict correctly. To exemplify, if the ratio is too low it means that the model missed too many positives. Being FN the number of false negatives, recall is defined in Eq. 11.

$$\text{recall} = \frac{TP}{TP + FN} \quad (11)$$

- **F1-score:** It combines precision and recall in a single value and allows to compare the performance between several models. F1-score is defined in Eq. 12.

$$F1 - \text{score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (12)$$

- **Support:** Number of predictions made for each class.
- **Accuracy:** It measures the ratio of cases that the model has succeeded, considering all the classes.

From this information we can mention that the model has a precision of 100% for the 'W', '9' and 'A' gestures, which means that in none of the predictions made has resulted in the 'W', '9' or 'A' gesture when they were not. This can be verified in the column 'It was ...' of Table II in which the 'W', '9' and 'A' gestures do not appear.

On the other hand, in the column 'Said ...' of Table II the 'S', '4', '5' and 'L' gestures do not appear, which means that they have a 100% of recall. This means that all predictions made for the 'S', '4', '5' and 'L' gestures have been accurate without having incorrect predictions.

In Fig. 11, the training metrics for each epoch were recorded, including accuracy and loss for training and validation sets. It is observed a correct learning of network parameters with the set of training with the passage of the epochs and with the validation set is observed that after the epoch number 6 does not improve performance significantly. It is worth mentioning that in this model was used a learning rate of 0.0005 for the optimizer Adam and that no extra KeyPoints were added on the hands.

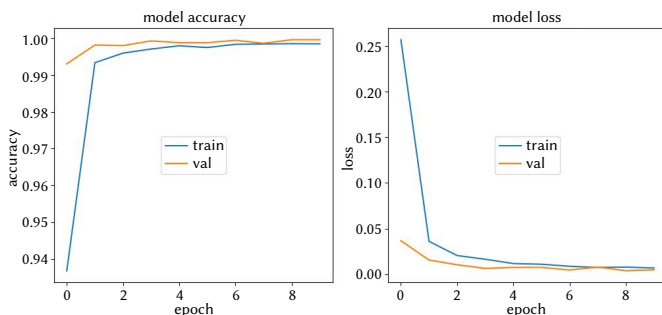


Fig. 11. Model accuracy and model loss for training number 1.

- **Model number 5:** For this model and in a comparative mode we will only analyze the metrics of Table IV and the graphs of Fig. 12. We can observe some minimal differences between precision and recall with respect to model number 1. However, we can use the f1-score metric to make a comparison with which we can indicate that the model number 5 has more erroneous predictions but is still very close to the performance of the previous model. Regarding the metrics during the learning process of the network, a similar behavior to the previous model is observed, where the performance does not improve considerably after the epoch number 10. For this model a learning rate of 0.0005 was used for Adam optimizer and 5 extra KeyPoints were added to each bone, making a total of 121 KeyPoints for each hand.

TABLE IV. METRICS FOR TRAINING NUMBER 5

gesture	precision	recall	f1-score	support
s	0.9977	1.0000	0.9989	870
1	0.9977	1.0000	0.9989	870
v	0.9988	0.9977	0.9983	870
w	1.0000	0.9954	0.9977	870
4	0.9966	0.9989	0.9977	870
5	0.9977	1.0000	0.9989	870
9	1.0000	1.0000	1.0000	870
L	0.9989	0.9989	0.9989	870
a	1.0000	0.9966	0.9983	870

accuracy = 0.9986 for 7830 predictions (870 each class)

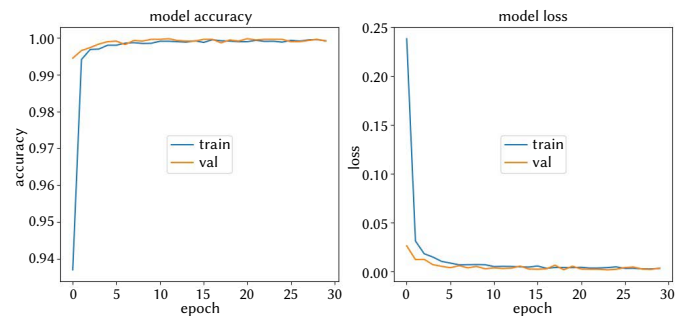


Fig. 12. Model accuracy and model loss for training number 5.

- **Model number 10:** In Table V a similar performance to the previous models is observed. No noticeable differences in the metrics during the learning (Fig. 13).

TABLE V. METRICS FOR TRAINING NUMBER 10

gesture	precision	recall	f1-score	support
s	0.9977	0.9989	0.9983	870
1	0.9977	1.0000	0.9989	870
v	0.9977	0.9977	0.9977	870
w	1.0000	0.9954	0.9977	870
4	0.9966	1.0000	0.9983	870
5	0.9989	0.9989	0.9989	870
9	1.0000	1.0000	1.0000	870
L	0.9989	0.9989	0.9989	870
a	0.9988	0.9966	0.9977	870

accuracy = 0.9985 for 7830 predictions (870 each class)

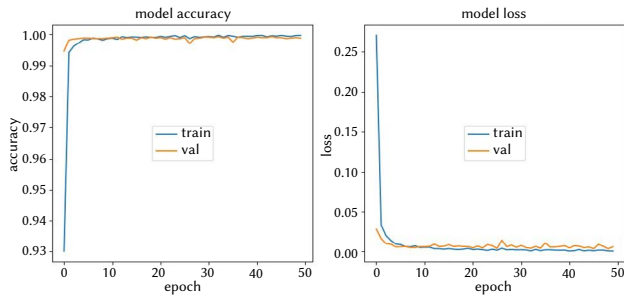


Fig. 13. Model accuracy and model loss for training number 10.

The results in Table I show a high performance of the proposed network. We detect the extra KeyPoints added on each bone would be of little importance, giving an indication that these extra KeyPoints do not provide significant information. We consider it would be important to include other type of information to the input data, such as the flexion angle at each joint and a number that identifies each KeyPoint. That is, if we look at Fig. 2 we can see that each KeyPoint has a number that identifies it and also on some KeyPoints is defined a flexion angle (except in the KeyPoints of the wrist and fingertips that do not have a defined angle). In this way, the input data could be defined as $(x, y, z, \text{number_kp}, \text{angle_joint})$.

C. Comparative Results

In order to compare the prediction accuracy of our model against other models, we have chosen our own test set (7,830 samples) and the external test set (4,500 samples). It is worth remembering that the own set is a random extraction of 20% of samples from our complete dataset (39,150 samples) and that the external test set is a collection of samples from third party works [27], [28], [29]. This set of external samples was made in order to obtain heterogeneous data, since they were extracted from images taken in other environments and by other people.

In addition to performing the predictions with our model (model number 1 in Table I) on the two test sets, we also use the PointNet model [13] trained with Adam optimizer with learning rate of 0.001 and 20 epochs, and also with a model created from ours, but without the initial transformation and normalization layer.

One can appreciate in these results the importance of the transformation and normalization layer that is initially applied. It provides a significant increase in accuracy when predictions are made with widely varying samples from different third party sources.

D. Comparison With ROC and AUC

The Receiver Operating Characteristic (ROC) is a measure of a classifier's predictive quality that compares and visualizes the tradeoff between True Positive Rate ($\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$) and False Positive Rate ($\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$). ROC curves are typically used in binary classification, but one of the ways it can be approached is by binarizing the output (per-class). A ROC curve displays the true positive rate on the Y axis and the false positive rate on the X axis. The ideal region is therefore the top-left corner of the plot, where false positives are zero and true positives are one. This leads to Area Under the Curve (AUC), which is a metric that relates false positives and true positives. The higher the AUC, in general, the better the model.

Fig. 14 presents the ROC curve of our model number 1 (from Table I) and shows the high success rate achieved in the predictions. All three models predict considerably well with our dataset, as shown in Table VI, and the ROC curves are very similar to the one presented in Fig. 14. We present the ROC curves of the three models performing the predictions with the external dataset. It can be observed that only our model with the normalization and transformation layer behaves in an acceptable performance. This is shown in Fig. 15, 16 and 17.

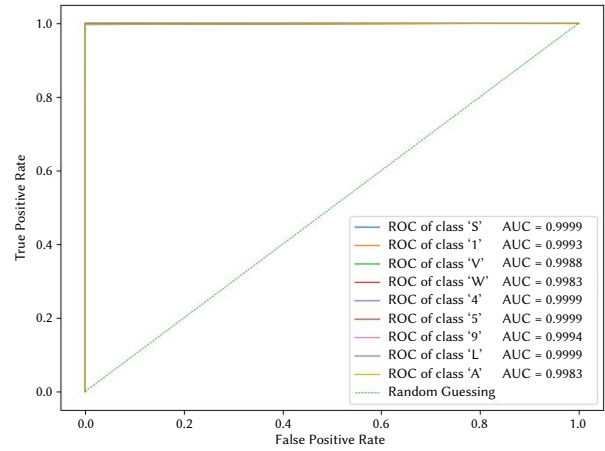


Fig. 14. ROC curves and AUC for our model with own dataset.

TABLE VI. COMPARISON OF MODELS

Model	Accuracy (our dataset)	Accuracy (external dataset)
Our model	7820 of 7830 99.87 %	4336 of 4500 96.36%
PointNet	7660 of 7830 97.82 %	2155 of 4500 47.89%
Our model without normalization	7760 of 7830 99.11 %	1371 of 4500 30.47 %

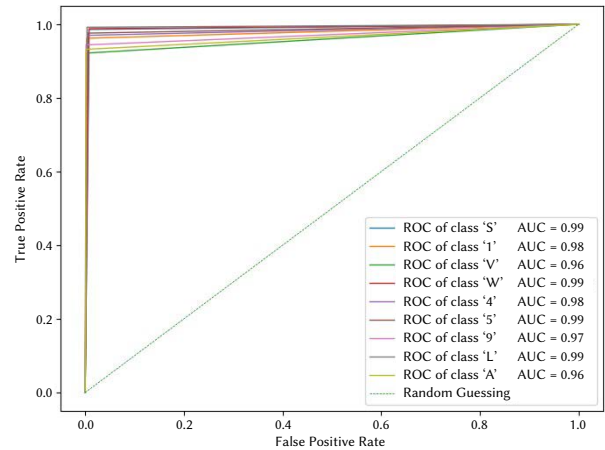


Fig. 15. ROC curves and AUC for our model with external dataset.

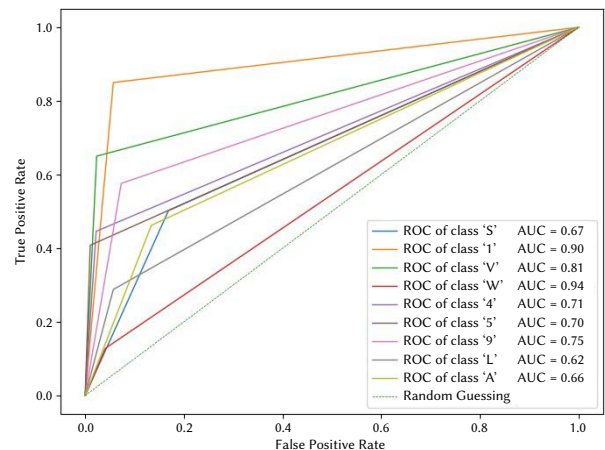


Fig. 16. ROC curves for PointNet model with external dataset.

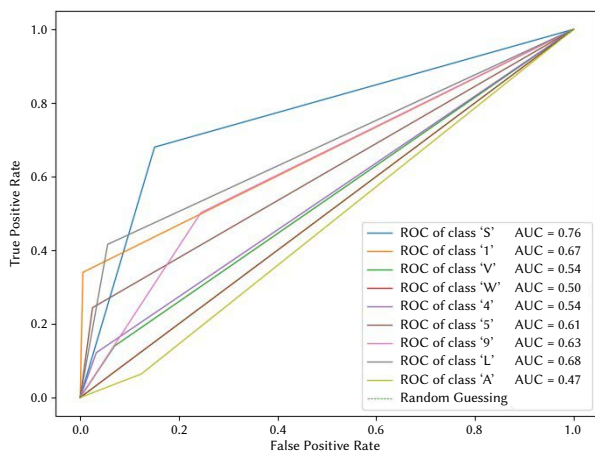


Fig. 17. ROC curves for our model without normalization.

V. CONCLUSION

In this work, we present a new network architecture for hand gesture recognition using point cloud. The study was focused on the cloud of 3D reference points obtained through a standard RGB camera. The new network (based on PointNet architecture) was trained with hand KeyPoints and thanks to a simple architecture with few hidden layers it is possible to work directly on the CPU.

The results show an accuracy of 99.87% in our hand gesture dataset. It is interesting to extend this study by including new gestures in order to have a wider variety of options for device control, and also to experiment with end users to detect those gestures that are more appropriate to perform certain control commands.

It is important to notice that the transformation and normalization layer allows us to maintain the good prediction performance of our model by using third-party datasets that contain a wide variety of users and physical spaces where samples are taken.

ACKNOWLEDGMENT

This work was funded by the EU ERDF and the Spanish Ministry of Economy and Competitiveness (MINECO) under AEI Project TIN2017-83964-R. <http://acg.ual.es/projects/cosmart/>

REFERENCES

- [1] M. Palieri, B. Morrell, A. Thakur, K. Ebadi, J. Nash, A. Chatterjee, C. Kanellakis, L. Carlone, C. Guaragnella, A. a. Aghamohammadi, "Locus: A multi-sensor lidar-centric solution for high-precision odometry and 3d mapping in real-time," *IEEE Robotics and Automation Letters*, pp. 1–1, 2020.
- [2] W. Zhang, D. Yang, "Lidar-based fast 3d stockpile modeling," in *2019 International Conference on Intelligent Computing, Automation and Systems (ICICAS)*, 2019, pp. 703–707.
- [3] S. Muhammad, G. Kim, "Visual object detection based lidar point cloud classification," in *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2020, pp. 438–440.
- [4] C. P. Hsu, B. Li, B. Solano-Rivas, A. R. Gohil, P. H. Chan, A. D. Moore, V. Donzella, "A review and perspective on optical phased array for automotive lidar," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 27, no. 1, pp. 1–16, 2021.
- [5] E. de Oliveira, E. W. Gonzalez, D. G. Trevisan, L. C. de Castro Salgado, "Investigating users' natural engagement with a 3d design approach in an egocentric vision scenario," in *2020 22nd Symposium on Virtual and Augmented Reality (SVR)*, 2020, pp. 74–82.
- [6] F. Zhang, V. Bazarevsky, A. Vakunov, G. Sung, C.-L. Chang, M. Grundmann, A. Tkachenka, "Mediapipe hands: On-device real-time hand tracking," in *CVPR Workshop on Computer Vision for Augmented and Virtual Reality*, June 2020.
- [7] Y. Kartynnik, A. Ablavatski, I. Grishchenko, M. Grundmann, "Real-time facial surface geometry from monocular video on mobile gpus," in *CVPR Workshop on Computer Vision for Augmented and Virtual Reality*, June 2019.
- [8] V. Bazarevsky, I. Grishchenko, K. Raveendran, M. Grundmann, F. Zhang, T. Zhu, "Blazepose: On-device real-time body pose tracking," in *CVPR Workshop on Computer Vision for Augmented and Virtual Reality*, June 2020.
- [9] S. Shi, X. Wang, H. Li, "Pointcnn: 3d object proposal generation and detection from point cloud," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 770–779.
- [10] Y. Zhou, O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4490–4499.
- [11] B. Li, "3d fully convolutional network for vehicle detection in point cloud," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 1513–1518, IEEE.
- [12] S. K. Arachchi, N. L. Hakim, H.-H. Hsu, S. V. Klimenko, T. K. Shih, "Real-time static and dynamic gesture recognition using mixed space features for 3d virtual world's interactions," in *2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, 2018, pp. 627–632, IEEE.
- [13] H. Seo, S. Joo, "Influence of preprocessing and augmentation on 3d point cloud classification based on a deep neural network: Pointnet," in *2020 20th International Conference on Control, Automation and Systems (ICCAS)*, 2020, pp. 895–899.
- [14] Zhirong Wu, S. Song, A. Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1912–1920.
- [15] Z. Li, W. Li, H. Liu, Y. Wang, G. Gui, "Optimized pointnet for 3d object classification," in *International Conference on Advanced Hybrid Information Processing*, 2019, pp. 271–278, Springer.
- [16] A. Jerjec, D. Bojanić, K. Bartol, T. Pribanić, T. Petković, S. Petrak, "On using pointnet architecture for human body segmentation," in *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, 2019, pp. 253–257, IEEE.
- [17] L. Ge, Y. Cai, J. Weng, J. Yuan, "Hand pointnet: 3d hand pose estimation using point sets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8417–8426.
- [18] Y. Yang, C. Feng, Y. Shen, D. Tian, "Foldingnet: Point cloud auto-encoder via deep grid deformation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 206–215.
- [19] Y. Yu, F. Li, Y. Zheng, M. Han, X. Le, "Clustering-enhanced pointcnn for point cloud classification learning," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–6.
- [20] C. R. Qi, L. Yi, H. Su, L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5099–5108, Curran Associates, Inc.
- [21] Y. Momma, W. Wang, E. Simo-Serra, S. Iizuka, R. Nakamura, H. Ishikawa, "P2net: A post-processing network for refining semantic segmentation of lidar point cloud based on consistency of consecutive frames," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp. 4110–4115.
- [22] D. Kulon, R. A. Guler, I. Kokkinos, M. M. Bronstein, S. Zafeiriou, "Weakly-supervised mesh-convolutional hand reconstruction in the wild," in *CVPR Workshop on Computer Vision for Augmented and Virtual Reality*, June 2020.
- [23] K. K. Verma, B. M. Singh, H. Mandoria, P. Chauhan, "Two-stage human activity recognition using 2d-convnet," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 6, no. 2, 2020.
- [24] M. Khari, A. K. Garg, R. G. Crespo, E. Verdú, "Gesture recognition of rgb and rgb-d static images using convolutional neural networks," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 5, no. 7, 2019.
- [25] M. Kim, J. Cho, S. Lee, Y. Jung, "Imu sensor-based hand gesture recognition

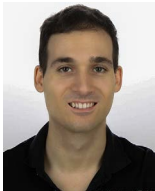
for human-machine interfaces,” *Sensors*, vol. 19, no. 18, p. 3827, 2019.

- [26] J.-H. Kim, G.-S. Hong, B.-G. Kim, D. P. Dogra, “deepgesture: Deep learning-based gesture recognition scheme using motion sensors,” *Displays*, vol. 55, pp. 38–45, 2018.
- [27] A. Thakur, “American Sign Language Dataset for Image Classification.” <https://www.kaggle.com/ayuraj/asl-dataset>, 2019. [Online; accessed 2-July-2021].
- [28] A. Memo, L. Minto, P. Zanuttigh, “Exploiting Silhouette Descriptors and Synthetic Data for Hand Gesture Recognition.” <https://lstm.dei.unipd.it/downloads/gesture/>, 2015. [Online; accessed 2-July-2021].
- [29] P. Bao, A. I. Maqueda, C. R. del Blanco, N. García, “Image database for tiny hand gesture recognition.” <https://sites.google.com/view/handgesturedb/home>, 2017. [Online; accessed 2-July-2021].
- [30] C. R. Qi, H. Su, K. Mo, L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.



César Osimani

César Osimani is an Associate Professor in the Applied Research & Development Center on IT at Universidad Blas Pascal, Argentina. He received the degree in Telecommunications Engineering and he is currently a PhD student at Universidad Nacional de Córdoba. He is engaged in the research on computer vision, pattern recognition, augmented reality and Human-Computer Interaction.



Juan Jesus Ojeda-Castelo

Computer Science Engineering at University of Almeria. Juan Jesus got a Master degree in Systems and Languages Programming about Computer Science by UNED and he is currently a Ph.D. student at University of Almeria. He is a collaborator of the Project titled Investigar en el uso didactico de la Kinect en el CEEE Princesa Sofia Almeria which is supported by Junta de Andalucía. He is very

interested in Human-Computer Interaction particularly Natural interaction, Computer Vision and Artificial Intelligence especially Deep Learning. The devices that he usually attempts to include in his personal projects are: Microsoft Kinect, Leap Motion, Intel RealSense, so far.



Jose Antonio Piedra-Fernandez

He received his PhD in Computer Science from the University of Almeria in 2005. Currently, he is an Assistant Professor at the Department of Informatics, the same University. He is member of the Research Applied Computing Group (TIC-211), Coordinator of the Master in Computer Engineering since 2014 and Director of the Quality Secretariat since 2017. He works closely with the

James Wang’s research group at The Pennsylvania State University. José Luis Labrandero Prize by the Spanish Association of Remote Sensing in 2007. He got a Patent in the field of recognition of cancer cells using a fuzzy robot vision system. Participates in various national, international and regional projects. He is author and co-author in several scientific publications, among journal articles, national and international book chapters, and publications in national and international congress proceedings. He is reviewer of scientific articles in several international JCR impact journals, such as IEEE Transactions on Geoscience and Remote Sensing or International Journal of Remote Sensing. His research area focuses on computer vision, artificial intelligence, natural interaction systems and serious games applied to the field of education and health.

ConvGRU-CNN: Spatiotemporal Deep Learning for Real-World Anomaly Detection in Video Surveillance System

Maryam Qasim Gandapur^{1*}, Elena Verdú²

¹ Department of Law, Shaheed Benazir Bhutto University, Dir (Upper), Khyber Pakhtunkhwa (Pakistan)

² Universidad Internacional de La Rioja, Logroño, La Rioja (Spain)

Received 14 April 2022 | Accepted 16 August 2022 | Published 30 May 2023



ABSTRACT

Video surveillance for real-world anomaly detection and prevention using deep learning is an important and difficult research area. It is imperative to detect and prevent anomalies to develop a nonviolent society. Real-world video surveillance cameras automate the detection of anomaly activities and enable the law enforcement systems for taking steps toward public safety. However, a human-monitored surveillance system is vulnerable to oversight anomaly activity. In this paper, an automated deep learning model is proposed in order to detect and prevent anomaly activities. The real-world video surveillance system is designed by implementing the ResNet-50, a Convolutional Neural Network (CNN) model, to extract the high-level features from input streams whereas temporal features are extracted by the Convolutional GRU (ConvGRU) from the ResNet-50 extracted features in the time-series dataset. The proposed deep learning video surveillance model (named ConvGRU-CNN) can efficiently detect anomaly activities. The UCF-Crime dataset is used to evaluate the proposed deep learning model. We classified normal and abnormal activities, thereby showing the ability of ConvGRU-CNN to find a correct category for each abnormal activity. With the UCF-Crime dataset for the video surveillance-based anomaly detection, ConvGRU-CNN achieved 82.22% accuracy. In addition, the proposed model outperformed the related deep learning models.

KEYWORDS

Anomaly Activities, Crime Detection, ConvGRU, Convolutional Neural Network (CNN), Deep Learning, Video Surveillance.

DOI: 10.9781/ijimai.2023.05.006

I. INTRODUCTION

WITH the growing public safety and security challenges, demand for increasing public safety monitoring through video surveillance cameras is also growing. Human-monitored surveillance systems can mine critical and helping cue from the patterns. This can help in detecting the abnormal activities for instant reaction [1]. However, owing to the human-monitored limitations, it is difficult to mine critical and helping cues [2]. Thus, an automated method to detect abnormal activities is critical. A sub-domain to understand behaviour from the video surveillance cameras is to detect anomaly activities [3]. The anomaly detection in the video surveillance is a crucial task and can face difficulties such as actions which do not tail definite patterns are termed as anomalies. Furthermore, actions are abnormal or normal in different situations indicating that a global abnormal activity can be a usual activity in certain situations such as gun club shooting. The shooting is usually an abnormal activity, but a normal activity in shooting clubs. Alternatively, some behavior is not

essentially abnormal, but might be anomalies in different situations [4]. According to some studies [5]-[6], abnormal actions ended at unusual locations and times.

Several kinds of abnormal activities are usually identified which include killing, looting, molestation, and intensive attacks. Killing is a deliberate action to kill a person. Looting is an action of stealing belongings from the people using extreme physical force and violence. Molestation is sexual exploitation of people (man, woman, and children) against their desire. This criminal activity is terrible and shows substantial consequences. Intensive attacks are illegal fights by one person against another to get something or to harm individuals [7]. Anomaly detection and prevention using deep learning is an attention-grabbing system. Many law enforcement organizations across the globe are experimenting deep learning systems to safeguard public safety. The anomaly activities are predictable and require high volume data processing, exposing the anomaly patterns which are informative for a law enforcement department. In some situations, an anomaly activity remains unreported because of external pressures from all verticals of society. For this reason, an intelligent security system is able to autonomously detect anomaly activities and supports in excluding manipulative activities by bypassing individuals and informing law enforcement departments. For example, there is a case

* Corresponding author.

E-mail address: maryam@sbbu.edu.pk

study of San-Francisco in USA and Natal in Brazil where anomaly activities have been predominant and monitored by the intelligent video surveillance systems [8].

Video camera surveillance is a key feature of monitoring systems [9]. Computer vision automates anomaly activity detection in videos by alarming a law reinforcement system when abnormal activity is observed to derive important information from recorded videos [10]-[11]. Various anomaly activities while entering or departing the public places require careful examination which might be important towards a pattern of anomaly activity [12]. In few situations, previous patterns can recognize malicious individuals in the recorded videos [13]-[14]. We can automate feedback activities when anomalies are observed to derive information from recorded videos using deep learning models [15]. According to deep learning perspective, the detection of the anomaly actions is divided into supervised, unsupervised, and semi-supervised learning models. In a single deep learning model, the model is trained on normal or abnormal activities [16]. On the other hand, both normal and abnormal activities are used to train deep learning models in multi-model learning setting [17]. Several studies took advantage of the supervised deep learning to detect anomaly activities in videos [18]-[25]. Many deep learning models including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long-Short Term Memory (LSTM), Gated Recurrent Units (GRUs), and Generative Adversarial Networks (GANs) are used for anomaly detection and prevention [26]-[28].

This study proposes anomaly activity detection in a multiple-learning perspective using a supervised deep learning model. Numerous abnormal activities in real-world are labeled as anomalies; but the focus of this study is on anomaly activities mentioned in the UCF-Crime dataset [29] which includes abnormal and violent behavior recorded by the video surveillance cameras in various public places. The proposed deep learning model for anomaly detection and prevention has implemented the ResNet-50 as a CNN model to extract high-level features from video frames. The CNN extracted features are fed to RNN model, ConvGRU, to learn temporal dependencies in the video dataset. The proposed deep learning model ConvGRU-CNN returns output indicating whether input videos include abnormal or normal behaviour. This ConvGRU-CNN model can reduce limitations of human-monitored video surveillance systems and can improve the accuracy of anomaly activity detection. In addition, ConvGRU-CNN can considerably improve the response-time. A compact neural model for anomaly detection is proposed by implementing a convolutional form of conventional GRU to learn temporal features videos. Alternative to the fully connected layer in GRU, a convolutional layer intensely reduces the parameters number. In addition, the incorporation of GRU further reduces the parameters when replaced with LSTM in ConvLSTM. There is 25% further reduction in parameters with ConvGRU. With UCF-Crime dataset, 13 classes of abnormal events are used to evaluate the proposed ConvGRU-CNN.

II. RELATED STUDIES ON VIDEO SURVEILLANCE SYSTEMS

The goal of anomaly detection system is to predict and prevent the abnormal (criminal) activities. Though, the conventional non-deep learning approaches are beneficial but they operate independently. Hence, a machine which is able to integrate the important aspects of conventional approaches would extremely be advantageous. A study [28] has compared the violent criminal patterns between the community's dataset and the real criminal statistical data by using Waikato Environment for Knowledge Analysis (WEKA) platform. Three models including the linear regression, additive regression, and decision stump are implemented. The linear regression on selecting the random samples in testing was able to handle randomness

showing a better detection among models and proved the success of deep learning in detecting the violent patterns and criminal trends.

A study [30] examined the anomaly detection in urban areas where anomaly has combined to grid size 200×250m and examined retrospectively. An ensemble model of logistic regression and neural network is proposed to detect anomaly. The results indicate that fortnightly predictions are improved remarkably as compared to monthly predictions. Anomaly activities are detected and examined in another work [31] using anomaly data of Vancouver for the last 15 years. A boosted decision tree and K-nearest neighbor (KNN) detected anomaly activities. A total of 560,000 records are examined and the anomaly activities are predicted with accuracy between 39% and 44%.

Another study [32] predicted anomaly statistics in Philadelphia to determine the trends of anomaly. Ordinal regression, KNN, logistic regression, and decision tree are trained with the datasets to get anomaly predictions. The models were able to determine the trend of anomaly activity with an accuracy of 69%. Data science models are implemented to detect the anomaly activities from the Chicago criminal dataset. Logistic regression, SVM/KNN classification, decision trees, random forest, and Bayesian models were examined and the most accurate model was selected for training. The KNN classification obtained the best accuracy of 78.7%.

A GUI-based deep learning model to predict the anomalies is presented in another study [33]. The results of supervised models are compared to predict anomalies. A feature-level data-fusion-based deep neural network (DNN) is proposed to predict anomaly with high accuracy by combining multi-model data from different domains with environmental context knowledge [34]. The data to train models (SVM, regression analysis, Kernel density estimation) was taken from online crime statistic database. SVM and KDE obtained 67.01% and 66.33% accuracies, whereas the proposed model obtained 84.25% accuracy. Another work [5] used previous crime locations to predict anomaly likely to happen in old locations. Bayesian neural networks, Levenberg Marquardt algorithm, and a scaled algorithm are implemented to examine and understand the data. The scaled algorithm showed the best results. The ANOVA verified that the scaled algorithm reduced crime rate by 78%, with 0.78 accuracy.

A framework to predict anomaly has been proposed [35] examining a dataset of formerly committed anomalies with their patterns. KNN and decision tree with adaptive boosting and random forest are implemented to boost the prediction accuracy. The records are divided into rare and frequent classes. The deep learning framework was trained with criminal activities recorded in a period of 12-years in San Francisco, USA. By applying oversampling and undersampling with random forest, 99.2% accuracy was achieved. Other studies have also achieved state of the art results for crime detection [36]-[44]. Table I summaries the previous work with achieved accuracies.

TABLE I. SUMMARY OF PREVIOUS WORK ON CRIME DETECTION

S. No.	Reference	Model	Achieved Accuracy
1	[36]	Decision Tree	59.15%
2	[37]	KNN	87.03%
3	[38]	Naive Bayes	87.00%
4	[39]	ARIMA	86.00%
5	[40]	Regression Model	72.00%
6	[41]	SVM	84.30%
7	[42]	Random Forrest	97.00%
8	[43]	E2E-VSDL	98.16%

Various CNN typed have been formulated such as the AlexNet, ResNets, VGG, Inceptions and their variants. Many studies combined these CNNs with a softmax layer [45], and morphological analysis [46] to detect anomaly. Besides CNN, other studies [47]-[48] proposed

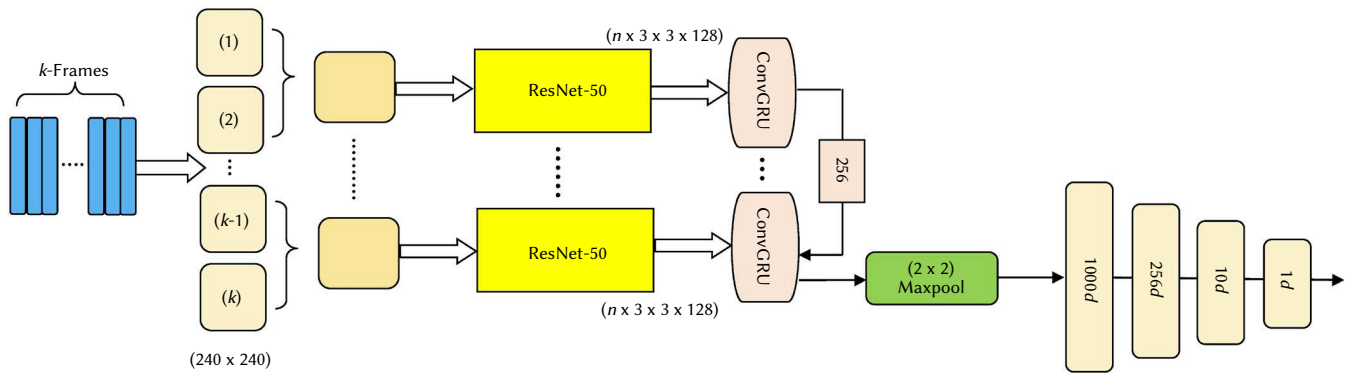


Fig. 1. The structure of the proposed ConvGRU-CNN.

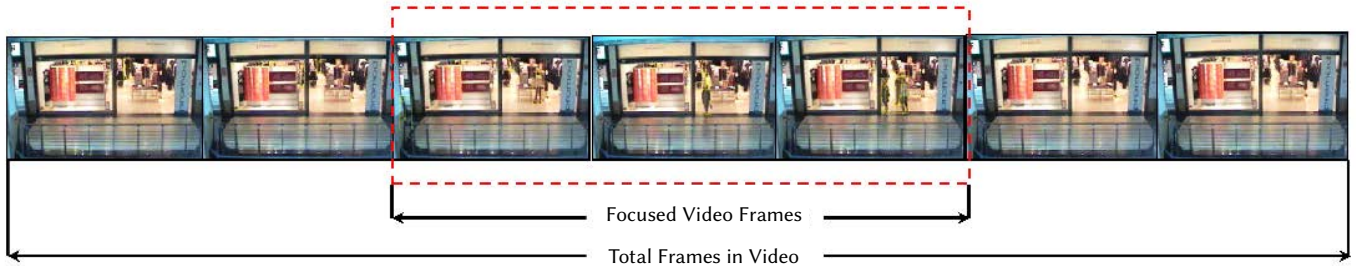


Fig. 2. Focused Bag frames extraction.

using autoencoders. Bayesian nonparametric [49] is also proposed to detect abnormal events in videos. Since surveillance camera feeds are sequential data, the LSTMs have gained attraction for anomaly detection. Encoder-decoder LSTM [50] is proposed in an unsupervised learning fashion. Spatiotemporal networks (STNs) are gaining popularity to learn spatial and temporal features [51] where RNNs and CNNs jointly extract spatiotemporal features for anomaly detection. ConvLSTM [52] is another model where a convolutional layer filters the output of CNNs before feeding a LSTM. Alternative to the fully connected layer in LSTM, a convolutional layer intensely reduces the parameters number. The GRUs further reduces the parameters if replaced with LSTM in ConvLSTM, obtaining a 25% reduction in the parameters. There are very limited studies implementing ConvGRU to detect anomaly in video streams.

III. PROPOSED ANOMALY DETECTION MODEL

Residual Networks (ResNets) are effective neural models to extract features in DNNs [53]. First, ResNet-50 is implemented to extract spatial features from the input video streams. In the next stage, the ConvGRU as RNN is used to extract the temporal dependencies in videos. The video streams are divided into sequences of k frames and fed to ResNet-50 as inputs. The outputs are further fed to ConvGRU. The spatiotemporal features are passed through maxpooling and fully connected layers to detect the anomaly.

A. Video Pre-Processing

Fig. 1 shows the proposed ConvGRU-CNN where input video is preprocessed and divided into fixed frames k with 30 frames/second. Therefore, for 60 sec video, the total number of frames is 1800. To consider the spatial movements for all input frames after selection, the difference between every frame and adjacent frames is calculated. Three categories from UCF-Crime dataset are selected. We split the exact time of the abnormal activity for each video and labelled them as Anomaly, such that, the remaining video is labelled as Normal. After that, the videos are divided into the same length. As a result, n frames are selected from k frames. Thus, only abnormal activities are focused.

Further, the normal activities are also selected from the same videos which include the anomaly activities. Except for actions, all other setting remains the same as in UCF-Crime dataset. Such arrangements help system in better detecting the anomaly activities. Full-length training videos result in a massive computational cost. Therefore, to understand the motion information in the recorded videos during training, we have considered a training framework over a defined set of frames, that is, focused bag which contains major information needed to understand the motions in the videos followed by block formation and selection. A set of frames composed of the activities in full length recorded video has been named as the focused bag and its extraction is shown in Fig. 2 where only a small part of the full-length recorded video is labelled as the suspicious/ criminal activity. Hence, L -frames out of M -frames are considered as a focused bag. This entire procedure is adopted and repeated for all recorded videos in the database thereby significantly minimized the training data by removing the redundant information.

B. ResNet-50 CNN Model

ResNets have shown excellent performance on many standard datasets such as ImageNet [16]. ResNets have many variants, such as, ResNet-18, ResNet-26, ResNet-50, ResNet-101, and ResNet-152. However, because of better performance and excellent architecture, ResNet-50 is instigated in ConvGRU-CNN. To avoid difficulties in labelling anomalies, Transfer Learning [54] is used in the model. As a result, ConvGRU-CNN is pre-trained on the ImageNet dataset, which includes 1000 sets of images. By executing ResNet-50 on ImageNet, the model parameters are initialized and updated thereby ready to execute on the preferred datasets. The input frame size is (240×240) allowing the ResNet-50 to process $(240 \times 240 \times 3)$ dimension data. After passing through the convolutional and pooling layers, a $4-d$ tensor $(n \times 1 \times 1 \times 2048)$ output is obtained from the Deep Residual Features (DRF), which is reshaped before fed to the ConvGRU filters. ResNet-50 structure is shown in Fig. 3 whereas the architecture is given in Table II. The ResNet50 output is reshaped into $(n \times 3 \times 3 \times 128)$ and is fed to ConvGRU layer. Since ResNet is not using for classification, the fully connected dense layer is not utilized.

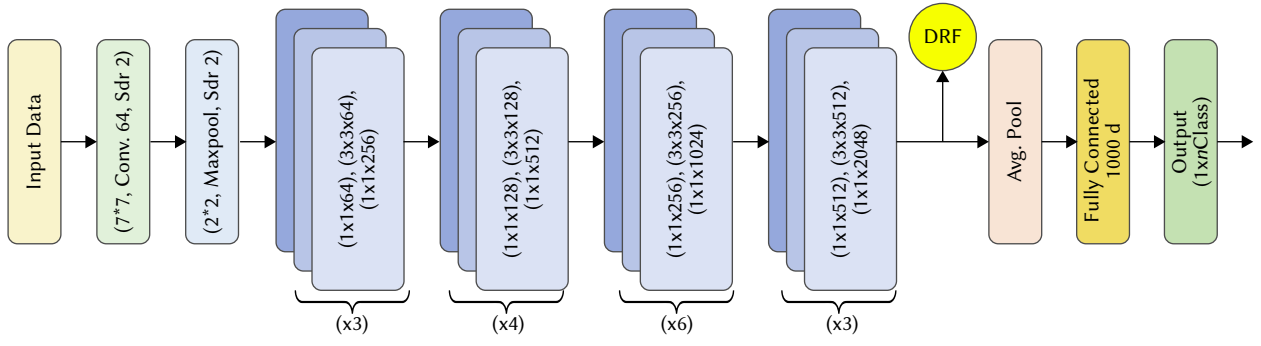


Fig. 3. ResNet-50 Structure.

TABLE II. RESNET-50 ARCHITECTURE

Layer Name	Output Size	50-Layers Model
Conv1	(112×112)	(7×7), 64, Stride 2 (3×3), Maxpool, Stride 2
Conv2	(56×56)	[(1×1, 64), (3×3, 64), (1×1, 256)]×3
Conv3	(28×28)	[(1×1, 128), (3×3, 128), (1×1, 512)]×4
Conv4	(14×14)	[(1×1, 256), (3×3, 256), (1×1, 1024)]×6
Conv5	(7×7)	[(1×1, 512), (3×3, 512), (1×1, 2048)]×3
	(1×1)	Average Pool, 1000d Fully Conn., Softmax
FLOPS		3.8 × 10 ⁹

C. ConvGRU Layer

The cell inputs, outputs, and states in GRUs are 1-d vectors; therefore, GRUs is unable to hold spatial relations between video pixels. As a result, GRUs is inappropriate for spatial sequence data [55]. In ConvGRU, due to the convolutional layers, cell states/inputs/outputs, and the spatial dimensions are 3-d tensors. Since the ConvGRU structure consists of convolutional gates, it can deal with spatial and temporal sequential data. The ConvGRU is a regular GRU but replaces the matrix multiplication with convolution operations. With convolution operations, the GRU can preserve spatial information. The formulation of the ConvGRU simply takes the standard linear GRU as:

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z) \quad (1)$$

$$r_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r) \quad (2)$$

$$c_t = \rho_g(W_c x_t + r * U_h h_{t-1} + b_c) \quad (3)$$

$$h_t = (z_t \odot h_{t-1} + (1 - r_t) \odot c_t) \quad (4)$$

Where W_z, W_r, W_c are weight matrices, b_z, b_r, b_c are biased terms, respectively, whereas x_t is input state [55]. By replacing the matrix multiplication with convolution operations (denoted as *), Eq. (1) - (4) became:

$$z_t = \sigma_g(W_z * x_t + U_z * h_{t-1} + b_z) \quad (5)$$

$$r_t = \sigma_g(W_r * x_t + U_r * h_{t-1} + b_r) \quad (6)$$

$$c_t = \rho_g(W_c * x_t + r * U_h * h_{t-1} + b_c) \quad (7)$$

$$h_t = (z_t \odot h_{t-1} + (1 - r_t) \odot c_t) \quad (8)$$

From the complexity viewpoint, GRU operates at 1-d vectors and after Hadamard product, the complexity increases due to large parameters size thereby the model is prone to overfitting. However, ConvGRU has a unique internal structure and requires fewer parameters which reduce the computational complexity of model. The video frames, after passing ResNet-50, feed the ConvGRU cell composed of 256 hidden states with (3×3) kernel size. The input to

ConvGRU is a 4-d tensor, ($n \times 256 \times 3 \times 3$) such that input at each time-step is (3×3) with 256 channels. The output of ConvGRU is maxpooled with (2×2) size and flattened to get a 1-d vector. The 1-d vector feed the fully-connected layers followed by batch normalization (BN) and ReLU activation. For binary classification (normal vs abnormal activity), sigmoid activation and binary cross entropy loss can be used after fully-connected layers. However, softmax with categorical cross entropy loss can be used for multi-class classification. The working flow of the proposed ConvGRU-CNN is given in Fig. 1.

IV. EXPERIMENTS

A. Dataset

In this paper, the proposed model is implemented on the UCF-Crime dataset [29] which includes abnormal, illegal and violent behaviour recorded by surveillance video cameras located in public places such as stores and streets. The UCF-Crime dataset is prepared from everyday actual events which is the key reason to select this dataset. Many studies have used handicraft datasets or particular datasets with the same backgrounds and environments (for example fighting and movies dataset), which is not according to our daily life. The UCF-Crime dataset is including lengthy surveillance video cameras feeds covering 13 different classes of anomaly events such as the Abuse, Arrest, Arson, Assault, Road Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism in addition to Normal events class. Fig. 4 shows example samples of the UCF-Crime dataset. For comparison with other related studies the training and testing data is arranged as (75%-25%) in experiments. Two variants of the UCF-Crime dataset including Ucfcrimes and Binary are used in the experiments where the Ucfcrimes contains 14 classes whereas Binary has 2 classes, one compiling the 13 abnormal activities and the normal one. The quantity of videos for each class from the Ucfcrimes and Binary datasets are given in Table III.

B. Model Settings and Model Selection

In the experiments the proposed model is applied by using ResNet-50 and ConvGRU, which are available in the Keras library. To tune the model, several hyperparameters are used to attain the best performance. Table IV shows the results of experiments with different types of weight initialization and optimizers. As a result, to initialize the ConvGRU-CNN weights, glorot-uniform (Xavier) is utilized whereas to optimize the model, RMSprop optimizer is imposed. The learning rate and number of epochs are fixed to 0.0001 and 100, respectively. However, early stop is applied when loss converges. The video sequence length is fixed to 20 frames. Since, focused bag is used for video frames. Table V provides a comparison of total video frames and frames in focus bag. In our experiments, we used different kinds of evaluations. During the first step, ConvGRU is tested with several CNN models such as InceptionV3, VGG19, ResNet-50, ResNet-101,

TABLE III. NUMBER OF VIDEOS FOR UCFCRIMES AND BINARY DATASETS

Anomalies	Videos (Ucfcrimers)	Videos (Binary)	Anomalies	Videos (Ucfcrimers)	Videos (Binary)
Abuse	50	50	Road Accident	50	150
Arrest	50	50	Robbery	50	150
Arson	50	50	Shooting	50	50
Assault	50	50	Shoplifting	50	50
Burglary	50	100	Stealing	50	100
Explosion	50	50	Vandalism	50	50
Fighting	50	50	Normal	50	950
Total	350	400	Total	350	1500

and ResNet-152, available in the Keras library. Table VI shows a comparison in terms of accuracy (in %). According to accuracy with less computational complexity, ResNet-50 was selected for integration with ConvGRU.

TABLE IV. HYPERPARAMETERS SETTINGS

Hyper Parameters	Tuning	Accuracy (%)
Weights Initialization	Glorot-Uniform (Xavier)	81.9%
Weights Initialization	Random-Uniform	80.2%
Weights Initialization	He-Uniform	80.2%
Optimizer	Adam	80.9%
Optimizer	RMSprop	81.3%

TABLE V. VIDEO FRAMES ANALYSIS (IN EXAMPLE ANOMALIES)

Anomaly	Total Video Frames	Focused Video Frames
Assault	130395	55023
Fighting	269255	132517
Shooting	157735	55427
Vandalism	157511	82163
Total	714896	156610

TABLE VI. CNN+CONVGRU COMPARISON

Hyper Parameters	Weights Initialization	Accuracy (%)
ResNet-50+ConvGRU	Glorot-Uniform (Xavier)	82.6%
ResNet-101+ConvGRU		RMSprop
ResNet-152+ConvGRU		86.3%
InceptionV3+ConvGRU	Adam	82.5%
VGG19+ConvGRU	RMSprop	89.3%

The evaluation measures are given by equations as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN} \quad (10)$$

$$\text{F1 - Score} = \frac{2 * TP}{2 * TP + FP + FN} \quad (11)$$

Where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

V. RESULTS AND DISCUSSIONS

First, the proposed ConvGRU-CNN model is examined by measuring the accuracy (Acc), precision (Prc), and F1-scores. Table VII provides the Acc, Prc, and F1 scores. It is clear from the Table VII that the proposed ConvGRU-CNN achieved significant metric scores for the 14 categories of the UCF-Crime dataset. The measuring results are averaged over the 14 types of activities, and the best accuracy

obtained is 82.22%. In addition, good precision and F1 are achieved with this considerable number of anomaly categories. The proposed model attained an encouraging average accuracy, precision, and F1-score of 82.88%, 82.89%, and 82.88%, respectively, at reducing the computational complexity. Therefore, an efficient model is proposed to analyze spatiotemporal features extracted from videos. Fig. 4 shows the detection of suspicious activity.

TABLE VII. MODEL EVALUATION IN TERMS OF ACC, PRC, AND F1 SCORES

Database	Accuracy	Precision	F1	AUC
Ucfcrimers	82.22%	83.13%	82.22%	82.65%
Binary	83.54%	85.65%	83.55%	82.77%
Average	82.88%	82.89%	82.88%	82.71%



Fig. 4. Detection of Normal and Suspicious Activities.

A. Comparison With Other Models

Limited literature on anomaly detection by using the UCF-Crime dataset is available. In the experiments, the proposed ConvGRU-CNN model is compared with other CNN models by measuring the Accuracy (Acc) and Area Under the Curve (AUC). Table VIII provides the AUC scores for the binary classification on the UCF-Crime dataset for the proposed model and other models for anomaly detection. The related models include support vector machine (SVM) [56], MIL [29], 3D-CNN [11], TSN [51], AutoEncd [48], SCL [57], CNN-RNN [58], and UGD-KM [59]. The categories for all the above mentioned abnormal events are considered as the Anomaly category whereas data with no abnormal events is considered as Normal. The testing classifier indicates the probabilities of correctly classified anomaly events. Table VIII shows that the proposed ConvGRU-CNN model outperformed the related benchmark models in anomaly detection. For example, AUC score is improved from 50.10% with SVM to 82.65% with ConvGRU-CNN and achieved 32.65% AUC gain. Similarly, AUC with AutoEncd is improved from 50.6% to 82.65% with large performance gain of 32.05%. Fig. 5 shows performance improvement over competing models. In comparison to 3D-CNN, the proposed ConvGRU-CNN improved the AUC from 81.05% to 82.65% whereas with MIL, the AUC is improved by 8.21%. The t-SNE results for normal and criminal activities are illustrated in Fig. 6.

TABLE VIII. MODEL COMPARISON IN TERMS OF AUC SCORES

Reference	Models	AUC (%)
(Erfani et al., 2016) [56]	SVM	50.10%
(Hasan et al., 2016) [48]	AutoEncd	50.60%
(Sultani et al., 2018) [29]	MIL (Loss with No Constraints)	74.44%
(Sultani et al., 2018) [29]	MIL (Loss with Constraints)	75.41%
(Zhong et al., 2019) [51]	TSN	78.08%
(Tran et al., 2015) [11]	3D-CNN	81.01%
(Vosta and Yow, 2022) [58]	CNN-RNN	81.77%
(Khan et al., 2018) [59]	UGD-KM	64.30%
Proposed	ConvGRU-CNN	82.65%

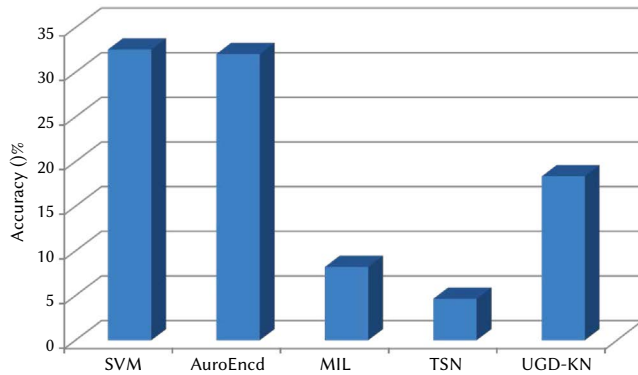


Fig. 5. percentage improvement over competing models.

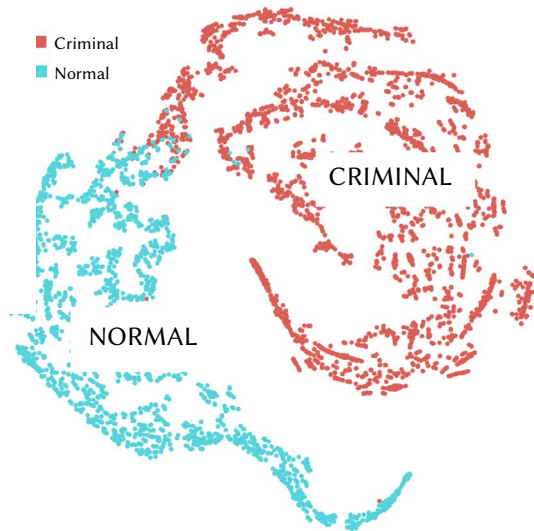


Fig. 6. t-SNE plots for normal and criminal activities.

To observe the efficiency of ConvGRU for crime detection, we have implemented ConvLSTM for the said task, and compared the accuracy, precision, F1, and AUC. This set of experiment was performed on the same experimental settings as done for ConvGRU. Table IX shows the results of the study. The results indicate that ConvLSTM underperforms in terms of accuracy, AUC, and computational cost, respectively.

TABLE IX. COMPARISON WITH LSTM IN TERMS OF ACCURACY, AUC AND COMPUTATIONAL COST

Database	Accuracy	AUC	Computational Cost
ConvLSTM	81.93%	81.98%	25% Less computational complexity with ConvGRU
ConvGRU	82.88%	82.71%	

VI. LAWS PREVENTING THE ANOMALY/CRIMINAL EVENTS

Anomaly events including criminal events and criminal intimidation has raised with hike in inflation in Pakistan and across the globe. It is imperative to devise new effective ways for preventing them. The conventional ways to detect the anomaly patterns are taking their part but technology has moved forward. A study [60] recommended deep learning models to the law enforcement agencies for predicting, detecting, and solving the anomaly activities at higher rates to reduce the crimes in society. So, it is recommended to use Artificial Intelligence (AI) to prevent the criminal events before their happening. The governments and other relevant institutes are responsible to prevent and reduce the criminal rate. As a result, the modern technology is one of the major solutions. To curtail this issue, law making authorities that is the legislature enacted Prevention of Electronic and other Crimes Act provided a detailed legal framework relating to different types of electronic crimes, procedures for the investigation and prosecution. Any act which is forbidden by the penal laws/ laws of the land amounts to criminal acts liable to be punished. With technological advancement, the already existing criminal patterns can be learnt to avoid future crimes and hence the punishment will become easy for law enforcement departments. To curtail heinous crimes, Serious Crimes Prevention Order is provided in Serious Crime Act 2007 in Pakistan. It is an adjudication order to safeguard public at large by preventing and restricting an individual participation in crimes.

VII. CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

This study proposes a novel deep learning framework by linking ResNet-50 and ConvGRU for detecting anomaly activities in the UCF-Crime dataset. Some anomalies took place in videos where persons cannot be seen such as car accidents. Besides, many anomaly events happen for few seconds and in a small length video (10 sec), most part of such videos shows a normal event. Regardless of stated limitations, the ConvGRU-CNN model outcores other models on the UCF-Crime dataset with 82.65% AUC and 82.88% accuracy. In addition to 14 classes of the UCF-Crime dataset, dividing the dataset into two major classes (Ucfrimes and Binary) shows improved results in terms of accuracy and AUC. The focused video frames extracted from the original videos of anomaly events have greatly improved the detection accuracy. Among other CNN models implemented for anomaly event, ResNet-50 [61] provides improved results when combined with ConvGRU. An excellent features extraction with ResNet-50 and ConvGRU significantly improved the performance measures. With ResNet-50+ConvGRU, the classifier efficiently detected the anomaly classes for Ucfrimes and Binary datasets. The experimental results demonstrate that ConvGRU-CNN performed better than other related models in terms of accuracy, precision, and AUC, yet we look to improve classification of all kinds of anomalies in the UCF-Crime dataset. One of the approaches is to add attention layers to the ConvGRU-CNN as future work. The attention layer is possible to integrate with CNN and/or ConvGRU. With this approach, the future model can be focused more precisely on the anomaly events in a video. Silent videos can be used to detect anomaly in terms of audio signals since most of the time only silent video is available. Therefore, if we incorporate audio signal synthesis, video surveillance can be made more effective [62].

REFERENCES

- [1] T. Hospedales, S. Gong, and T. Xiang, "Video behaviour mining using a dynamic topic model," *International journal of computer vision*, vol. 98, no. 3, pp. 303-323, 2012.
- [2] M. Cristani, R. Raghavendra, A. Del Bue, and V. Murino, "Human

- behavior analysis in video surveillance: A social signal processing perspective," *Neurocomputing*, vol. 100, pp. 86-97, 2013.
- [3] B. Tian, B.T. Morris, M. Tang, Y. Liu, Y. Yao, C. Gou, ... and S. Tang, "Hierarchical and networked vehicle surveillance in ITS: a survey," *IEEE transactions on intelligent transportation systems*, vol. 16, no. 2, pp. 557-580, 2017.
- [4] J. Yu, K.C. Yow, and M. Jeon, "Joint representation learning of appearance and motion for abnormal event detection," *Machine Vision and Applications*, vol. 29, no. 7, pp. 1157-1170, 2018.
- [5] S. R. Bandekar and C. Vijayalakshmi, "Design and analysis of machine learning algorithms for the reduction of crime rates in India," *Procedia Computer Science*, vol. 172, no. 122-127, 2020.
- [6] J. Varadarajan and J. M. Odobez, "Topic models for scene analysis and abnormality detection," in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, IEEE, 2009*, pp. 1338-1345.
- [7] C. Tabedzki, A. Thirumalaiswamy, P. van Vliet, S. Agarwal and S. Sun, "Yo home to Bel-Air: predicting crime on the streets of Philadelphia," University of Pennsylvania, CIS, 520, 2018.
- [8] K. A. Joshi and D.G. Thakore, "A survey on moving object detection and tracking in video surveillance system," *International Journal of Soft Computing and Engineering*, vol. 2, no. 3, pp. 44-48, 2012.
- [9] R. Socha and B. Kogut, "Urban video surveillance as a tool to improve security in public spaces," *Sustainability*, vol. 12, no. 15, 6210, 2020.
- [10] A. Selvaraj, J. Selvaraj, S. Maruthaiappan, G.C. Babu and P.M. Kumar, "L1 norm based pedestrian detection using video analytics technique," *Computational Intelligence*, vol. 36, no. 4, pp. 1569-1579, 2020.
- [11] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489-4497.
- [12] L. Alkanhal, D. Alotaibi, N. Albrahim, S. Alrayes, G. Alshemali and O. Bchir, "Super-resolution using deep learning to support person identification in surveillance video," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 7, 2020.
- [13] J. Athanesious, V. Srinivasan, V. Vijayakumar, S. Christobel and S.C. Sethuraman, "Detecting abnormal events in traffic video surveillance using superior orientation optical flow feature," *IET Image processing*, vol. 14, no. 9, pp. 1881-1891, 2020.
- [14] H. Zhang, P. Li, Z. Du and W. Dou, "Risk entropy modeling of surveillance camera for public security application," *IEEE Access*, vol. 8, pp. 45343-45355, 2020.
- [15] B.S. Harish and S.A. Kumar, "Anomaly based Intrusion Detection using Modified Fuzzy Clustering," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 6, pp. 54-60, 2017.
- [16] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, IEEE, 2009, pp. 248-255.
- [17] A.A. Sodemann, M.P. Ross and B.J. Borghetti, "A review of anomaly detection in automated surveillance," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1257-1272, 2012.
- [18] I.V. Pustokhina, D.A. Pustokhin, T. Vaiyapuri, D. Gupta, S. Kumar and K. Shankar, "An automated deep learning based anomaly detection in pedestrian walkways for vulnerable road users safety," *Safety science*, vol. 142, 105356, 2021.
- [19] R. Nawaratne, D. Alahakoon, D. De Silva and X. Yu, "Spatiotemporal anomaly detection using deep learning for real-time video surveillance," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 393-402, 2019.
- [20] M. Á. López, J.M. Lombardo, M. López, C.M. Alba, S. Velasco, M.A. Braojos and M. Fuentes-García, "Intelligent Detection and Recovery from Cyberattacks for Small and Medium-Sized Enterprises," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 3, 2020.
- [21] K. Rezaee, S.M. Rezakhani, M.R. Khosravi and M.K. Moghimi, "A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance," *Personal and Ubiquitous Computing*, pp. 1-17, 2021.
- [22] F. Rezaei and M. Yazdi, "A New Semantic and Statistical Distance-Based Anomaly Detection in Crowd Video Surveillance," *Wireless Communications and Mobile Computing*, vol. 2021, 5513582, 2021.
- [23] W. Ullah, A. Ullah, I.U. Haq, K. Muhammad, M. Sajjad and S.W. Baik, "CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks," *Multimedia Tools and Applications*, vol. 80, no. 11, pp. 16979-16995, 2021.
- [24] W. Ullah, A. Ullah, T. Hussain, Z.A. Khan and S.W. Baik, "An Efficient Anomaly Recognition Framework Using an Attention Residual LSTM in Surveillance Videos," *Sensors*, vol. 21, no. 8, 2811, 2021.
- [25] Y. Luo, Y. Xiao, L. Cheng, G. Peng and D. Yao, "Deep learning-based anomaly detection in cyber-physical systems: Progress and opportunities," *ACM Computing Surveys (CSUR)*, vol. 54, no. 5, pp. 1-36, 2021.
- [26] K.K. Santhosh, D.P. Dogra, P.P. Roy and A. Mitra, "Vehicular Trajectory Classification and Traffic Anomaly Detection in Videos Using a Hybrid CNN-VAE Architecture," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 11891-11902, 2021.
- [27] H. Fanta, Z. Shao and L. Ma, "SiTGRU: single-tunnelled gated recurrent unit for abnormality detection," *Information Sciences*, vol. 524, pp. 15-32, 2020.
- [28] W. Shin, S.J. Bu and S.B. Cho, "3D-convolutional neural network with generative adversarial network and autoencoder for robust anomaly detection in video surveillance," *International Journal of Neural Systems*, vol. 30, no. 6, 2050034, 2020.
- [29] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479-6488.
- [30] A. Rummens, W. Hardyns and L. Pauwels, "The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context," *Applied geography*, vol. 86, pp. 255-261, 2017.
- [31] S. Kim, P. Joshi, P.S. Kalsi and P. Taheri, "Crime analysis through machine learning," in *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, IEEE, 2018, pp. 415-420.
- [32] V. Tsakanikas and T. Dagiuklas, "Video surveillance systems-current status and future trends," *Computers & Electrical Engineering*, vol. 70, pp. 736-753, 2018.
- [33] S. Prithi, S. Aravindan, E. Anusuya and A.M. Kumar, "GUI based prediction of crime rate using machine learning approach," *International Journal of Computer Science and Mobile Computing*, vol. 9, no. 3, pp. 221-229, 2020.
- [34] H.W. Kang and H.B. Kang, "Prediction of crime occurrence from multimodal data using deep learning," *PLOS ONE*, vol. 12, no. 4, e0176244, 2017.
- [35] S. Hossain, A. Abtahee, I. Kashem, M.M. Hoque and I.H. Sarker, "Crime prediction using spatio-temporal data," in *International Conference on Computing Science, Communication and Security*, Springer, Singapore, 2020, pp. 277-289.
- [36] G.N. Obuandike, I. Audu and A. John, "Analytical study of some selected classification algorithms in WEKA using real crime data," *International Journal of Advanced Research in Artificial Intelligence*, vol. 4, no. 12, 2015.
- [37] C. C. Sun, C. Yao, X. Li and K. Lee, "Detecting Crime Types Using Classification Algorithms," *Journal of Digital Information Management*, vol. 12, no. 5, pp. 321-327, 2014.
- [38] M. Jangra and S. Kalsi, "Crime analysis for multistate network using naive Bayes classifier," *International Journal of Computer Science and Mobile Computing*, vol. 8, no. 6, pp. 134-143, 2019.
- [39] F. Vanhoenshoven, G. Nápoles, S. Bielen and K. Vanhoof, "Fuzzy cognitive maps employing ARIMA components for time series forecasting," in *International Conference on Intelligent Decision Technologies*, Springer, Cham, 2017, pp. 255-264.
- [40] W. Gorr, A. Olligschlaeger and Y. Thompson, "Assessment of crime forecasting accuracy for deployment of police," *International journal of forecasting*, 743-754, 2000.
- [41] C.H. Yu, M.W. Ward, M. Morabito and W. Ding, "Crime forecasting using data mining techniques," in *2011 IEEE 11th international conference on data mining workshops*, IEEE, 2011, pp. 779-786.
- [42] L.G. Alves, H.V. Ribeiro and F.A. Rodrigues, "Crime prediction through urban metrics and statistical learning," *Physica A: Statistical Mechanics and its Applications*, vol. 505, pp. 435-443, 2018.
- [43] M.Q. Gandapur, "E2E-VSDL: End-to-end video surveillance-based deep learning model to detect and prevent criminal activities," *Image and Vision Computing*, vol.123, 104467, 2022.

- [44] M. Adimoolam, S. Mohan, A. John and G. Srivastava, "A Novel Technique to Detect and Track Multiple Objects in Dynamic Video Surveillance Systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, 2022.
- [45] P. Christiansen, L.N. Nielsen, K.A. Steen, R.N. Jørgensen and H. Karstoft, "DeepAnomaly: Combining background subtraction and deep learning for detecting obstacles and anomalies in an agricultural field," *Sensors*, vol. 16, no. 11, 1904, 2016.
- [46] L. Dong, Y. Zhang, C. Wen and H. Wu, "Camera anomaly detection based on morphological analysis and deep learning," in *2016 IEEE International Conference on Digital Signal Processing (DSP)*, IEEE, 2016, pp. 266-270.
- [47] D. Xu, E. Ricci, Y. Yan, J. Song and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," 2015, arXiv preprint arXiv:1510.01553.
- [48] M. Hasan, J. Choi, J. Neumann, A.K. Roy-Chowdhury and L.S. Davis, "Learning temporal regularity in video sequences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 733-742.
- [49] V. Nguyen, D. Phung, D.S. Pham and S. Venkatesh, "Bayesian nonparametric approaches to abnormality detection in video surveillance," *Annals of Data Science*, vol. 2, no. 1, pp. 21-41, 2015.
- [50] T. Ergen and S.S. Kozat, "Unsupervised anomaly detection with LSTM neural networks," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 8, pp. 3127-3141, 2019.
- [51] J.X. Zhong, N. Li, W. Kong, S. Liu, T.H. Li and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1237-1246.
- [52] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, IEEE, 2017, pp. 1-6.
- [53] K. Zhang, M. Sun, T.X. Han, X. Yuan, L. Guo and T. Liu, "Residual networks of residual networks: Multilevel residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 6, pp. 1303-1314, 2017.
- [54] Y. Wu, Q. Wu, N. Dey and S. Sherratt, "Learning models for semantic classification of insufficient plantar pressure images," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 1, pp. 51-61, 2020.
- [55] M.G. Huddar, S.S. Sannakki and V.S. Rajpurohit, "Attention-based Multimodal Sentiment Analysis and Emotion Detection in Conversation using RNN," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 112-121, 2021.
- [56] S.M. Erfani, S. Rajasegarar, S. Karunasekera and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," *Pattern Recognition*, vol. 58, pp. 121-134, 2016.
- [57] C. Lu, J. Shi and J. Jia, "Abnormal event detection at 150 fps in matlab," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2720-2727.
- [58] S. Vosta and K.C. Yow, "A CNN-RNN Combined Structure for Real-World Violence Detection in Surveillance Cameras," *Applied Sciences*, vol. 12, no. 3, 1021, 2022.
- [59] M.U.K. Khan, H.S. Park and C.M. Kyung, "Rejecting motion outliers for efficient crowd anomaly detection," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 2, pp. 541-556, 2018.
- [60] Hosseinzadeh, M., Rahmani, A. M., Vo, B., Bidaki, M., Masdari, M., & Zangakani, M. "Improving security using SVM-based anomaly detection: issues and challenges," *Soft Computing*, vol. 25, 3195-3223.
- [61] A.A. Alvarez and F. Gómez, "Motivic Pattern Classification of Music Audio Signals Combining Residual and LSTM Networks," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, 2021.
- [62] N. Saleem, J. Gao, M. Irfan, E. Verdu and J.P. Fuente, "E2E-V2SResNet: Deep residual convolutional neural networks for end-to-end video driven speech synthesis," *Image and Vision Computing*, vol. 119, 104389, 2022.



Maryam Qasim Gandapur

Maryam Qasim received her LLB and LLM degrees from Khyber Law College, University of Peshawar in 2013 and 2015, respectively. She is preparing for PhD studies. She is currently an Assistant Professor at Department of Law, Shaheed Benazir Bhutto University, Dir, Khyber Pakhtunkhwa, Pakistan. She has been attached with academia for many years. Her research has focused on Corporate Law, Comparative Human Rights Law, intelligent systems for Law enforcement, and deep learning systems for Security.



Elena Verdú

Elena Verdú received her master's and Ph.D. degrees in telecommunications engineering from the University of Valladolid, Spain, in 1999 and 2010, respectively. She is currently an Associate Professor at Universidad Internacional de La Rioja (UNIR) and member of the Research Group "Data Driven Science" of UNIR. For more than 15 years, she has worked on research projects at both national and European levels. Her research has focused on e learning technologies, intelligent tutoring systems, competitive learning systems, accessibility, data mining and expert systems.

An Empirical Evaluation of Machine Learning Techniques for Crop Prediction

G. Mariammal¹, A. Suruliandi², S.P. Raja³, E. Poongothai⁴ *

¹ Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai - 600 062, Tamilnadu, (India)

² Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli – 627012, Tamilnadu, (India)

³ School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, (India)

⁴ Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamilnadu, (India)

Received 5 June 2021 | Accepted 11 April 2022 | Published 23 December 2022



ABSTRACT

Agriculture is the primary source driving the economic growth of every country worldwide. Crop prediction, which is critical to agriculture, depends on the soil and environment. Nutrient levels differ from area to area and greatly influence in crop cultivation. Earlier, the tasks of crop forecast and cultivation were undertaken by farmers themselves. Today, however, crop prediction is determined by climatic variations. This is where machine learning algorithms step in to identify the most relevant crop for cultivation. This research undertakes an empirical analysis using the bagging, random forest, support vector machine, decision tree, Naïve Bayes and k-nearest neighbor classifiers to predict the most appropriate cultivable crop for certain areas, based on environment and soil traits. Further, the suitability of the classifiers is examined using a GitHub prisoners' dataset. The experimental results of all the classification techniques were assessed to show that the ensemble outclassed the rest with respect to every performance metric.

KEYWORDS

Classification, Crop Prediction, Environmental Characteristics, Machine Learning, Soil Characteristics.

DOI: 10.9781/ijimai.2022.12.004

I. INTRODUCTION

AGRICULTURE is key to the development of human civilization, with farming playing a critical role in the process. Crop cultivation varies across areas, with each possessing unique soil, climatic and geographic characteristics. Soil is central to crop cultivation, and nutrients namely potassium, nitrogen, and phosphorus impact yield. Geography and climatic conditions, including the seasons, soil types, rainfall, and temperature also greatly influence in crop prediction. Based on these factors, the most suitable cultivable crop is predicted using several Machine Learning (ML) [1] techniques. Classification is fundamental to machine learning, for which it trains the system to obtain results using the given data. The supervised, unsupervised and reinforcement learning types of classification techniques are used in prediction. This research evaluates the performance of supervised learning techniques such as bagging, random forest (RF), support vector machine (SVM), decision tree (DT), Naïve Bayes (NB) and k-nearest neighbor (kNN) to predict a relevant crop for classification, using a GitHub prisoners' dataset. This work identifies the best classifier for the forecasting process.

A. Related Work

Several papers that illustrate key features of common ML models are discussed in this section.

Soil characteristics alone are used to predict a suitable crop for cultivation [1]. Belson et al. [2] described the DT classification model as a tree structure, with leaf nodes representing the final decision made after the top-to-bottom path is established. The most efficient techniques used in the literature survey include the Gaussian mixture, the Chi-square Automatic Interaction Detector (CHAID), classification and regression trees, and the Bayesian network, presented by Duda et al. [3], Kass et al. [4], Breiman et al. [5], and Neapolitan [6], respectively. The NB classification technique, built on the Bayesian theorem, produces accurate forecast results that are easy to train and classify. Kohonen [7] and Atkeson et al. [8] discussed memory-based models and constructed hypotheses directly from the available data. However, information overload can increase their complexity. Data mining techniques with applications in agriculture include the K-means algorithm to forecast atmospheric emissions, the kNN to model daily precipitation and miscellaneous climatic variables, and the SVM to analyze possible adjustments to the weather. Bayesian models such as the NB, Gaussian NB and multinomial NB used in the prediction process were explained [9] - [11].

Ensemble learning (EL) models enhance the prediction process by constructing a prediction model using single base learners. Ensemble techniques such as bagging, boosting and the AdaBoost algorithms

* Corresponding author.

E-mail addresses: suba.g1212@mail.com (G. Mariammal), suruliandi@yahoo.com (A. Suruliandi), avemariaraja@gmail.com (S.P. Raja), poongothai.rp@gmail.com (E. Poongothai)

were discussed and implemented [12] - [14]. The widely used SVM technique was improved through the use of the kernel trick for prediction [15], [16]. Breiman [17] proposed an RF technique, which is an ensemble model, and combined it with several DTs to constitute a single tree for a prediction model. The results are obtained after a comparison of all the trees in the forest and a final decision is made, based on the voting method. Suykens et al. [18] presented the minimum squares SVM, and Galvao et al. [19], the successive vector support algorithm. The proximity of data points is shown in the decision surface, that is, hyperplane support vectors. The data in the hyperplane are linearly separated by the total distance in the SVM. Babu et al. [20] discussed the application of artificial intelligence and ML algorithms in crop prediction. Designing an expert framework for crop cultivation calls for the services of computer engineers to model it, agricultural scientists to program it, and the know-how of experts in the field to back it up. Veenadhari et al. [21] described the role of data mining in agriculture. The most suitable crop for cultivation was predicted with 95% accuracy, based on climatic conditions as a major feature.

Monali et al. [22] posited a prediction system that categorizes soil types and predicts crop yields using the NB and kNN methods. Jeong et al. [23] explained the ability of the RF to predict crop yield responses to global and regional weather as well as biophysical variables in wheat, maize and potato. Sellam et al. [24] discussed crop yield prediction, which is primarily dependent on environmental characteristics, using regression analysis and linear regression. In their work, Pudumalar et al. [25] proposed a new ensemble model using the random tree, CHAID, kNN and NB to recommend crops for specific zones.

Zala [26] described bagging as a meta-algorithm that complements the power and precision of the ML technique used in mathematical classification and regression. It also eliminates variations and averts overfitting. Balducci et al. [27] described the DT as a predictive model and tested it at every level requiring decisions to be made. The levels depend on the request and outcome of the decision-making process. Jahan [28] averred that the NB is vulnerable to insignificant characteristics. Given its solid foundation, it manages both confidential and streaming data with ease. Priya et al. [29] used real-time Tamil Nadu facts to predict crop yields the usage of the RF method. Suresh et al. [30] examined soil profiles in conjunction with Global Positioning System-based technologies. The K-means and modified kNN are implemented to predict crop yields in Tamil Nadu.

B. Motivation and Justification

Crop prediction, which is critical to agriculture, employs machine learning algorithms for the purpose. Classification is central to machine learning [40]-[42]. It helps to learn the system for forecasting a relevant cultivable crop. Classifiers are divided into two sub-categories, single learner and ensemble learners. Thus motivated, various supervised classifiers are examined for the prediction process. Though the literature analysis makes it evident that the ensemble model offers better predictions, much of the research has, however, tended to use single learners for crop prediction. An ensemble model, which helps improve the prediction rate, is constructed using single learners. Thus justified, the efficiency of the ensemble model is examined with a crop dataset and a GitHub prisoners' dataset, using different performance metrics. The performance of the ensemble bagging model is evaluated with existing classification algorithms such as the RF, SVM, DT, NB and kNN for the prediction process.

C. Contributions

The significance contributions of this research are given below:

- The literature survey shows that much of the earlier work has examined either soil or environmental factors to predict crop

cultivation. This work, on the other hand, undertakes crop prediction by examining both.

- A real-time dataset composed from the Sankarankovil Agriculture Department of Tenkasi District in the state of Tamil Nadu in India is used for the prediction process.
- The primary goal of this work is to predict an appropriate classifier for all sort of datasets.
- Further, the classifier performance is examined by the various k-fold and data splitting methods.

D. Outline of the Work,

Fig. 1 illustrates the comprehensive process of this work. Input data is fed into pre-processing step. In pre-processing, missing values in the dataset are identified to eliminate the redundant values. This is used to handle the imbalanced data which was done by mean imputation method. It improves the prediction performance of classifiers and accuracy rate has been increased after pre-processing stage. After that, the dataset is broken down into training and testing. Classifiers are well trained to predict the target class with the help of all training samples. The learned classifier is validated with the unknown samples from the testing dataset. The learned classifier helps to forecast the target class of the given dataset. Finally, the predicted result is examined by various performance metrics.

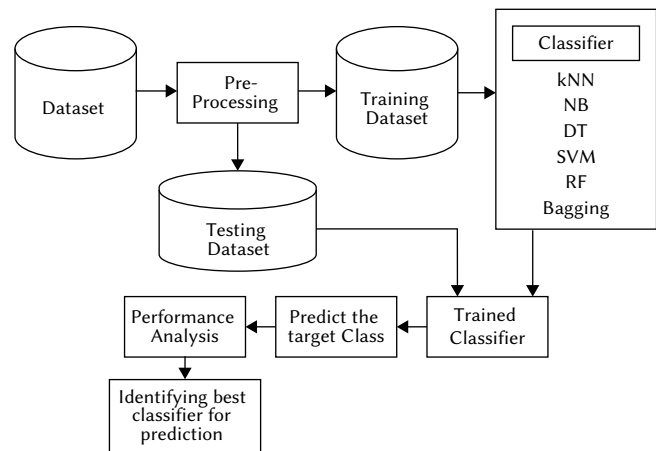


Fig. 1. Outline of the Work.

E. Organization of the Paper

The remainder of the article is organized in the following way: Section II describes the methodology for crop prediction. Section III illustrates the experimental results and final section concludes this work.

II. METHODOLOGY

A. Classification

Classification is indispensable to machine learning, given that it predicts the outcome of the process. This work evaluates the performance of the existing bagging, RF, SVM, DT, NB and kNN classifiers, using a crop dataset and a GitHub prisoners' dataset.

1. K Nearest Neighbor (KNN)

The kNN is not a complex algorithm that classifies new instances established on positive similarity measures [25]. The similarity degree is calculated by distance measures which include Euclidean distance, Manhattan distance, and many others [31]. In the kNN, feature vectors are stored in the training phase of the algorithm. The kNN technique

finds the similarity between unknown classes with known instances. The class labeling of training instances and unlabeled vectors are classified by way of assigning the most common label of the closest training samples. In the iterative classification, k is a parameter set by the user [27]. Algorithm 1 gives the pseudo code of the kNN classifier.

Algorithm 1. The pseudo code for the kNN

*Input: s, C, d where s is unknown sample from dataset C ;
 d is the distance
 Output: class of s
 for $(s', c) \in D$ do
 Compute the distance $d(s', s)$
 end for
 Order the $|C|$ distances by increasing the sequence
 Calculate the number of hits for each class c_i among the kNN
 Assign s to the highest class*

2. Naïve Bayes (NB)

In order to construct classifiers, the NB method offers class labels to problem instances [25] which is entrenched the theorem of Bayes' [28]. NB is one of the most effective classifiers and it predicts the outcome based on the probability of an instance. It deems the value of a separate variable to be independent of the value of any other given quality in the class variable. [25]. It is utilized for both binary and multi class classification problems. Algorithm 2 gives the pseudo code of the NB classifier [32].

Algorithm 2. The pseudo code for the NB

*Input: $C \rightarrow$ Dataset, $T_1 \rightarrow$ Training dataset, $P = (p_1, p_2, \dots, p_n)$ //
 value of the predictor variable in testing dataset
 Output: Predicted class*

Step:

1. Read all training data T_1
2. The mean and standard deviation of the predictive variable in each category are calculated
3. Repeat
 Compute the likelihood of p_i using the gauss density equation in each class
 Until the likelihood of all predictor variables (p_1, p_2, \dots, p_n) has been computed
4. Compute the probability of each class
5. Obtain the highest likelihood

3. Decision Tree (DT)

The DT is a single tree predictive model, and it is used for both classification and regression problems. DT is like a tree structure method [27]. Decision nodes and leaves make up a tree [31]. Each internal node shows an input variable, and each leaf node shows class prediction. It works supported a top-down approach by selecting a worth for the feature at every stop that best splits a collection of things [27], looking on the applying and decision-making outcome. DT algorithms contain the CART, C4.5 and ID3. In this work CART technique is used for implementing the DT algorithm which stands for classification and regression tree. Algorithm 3 gives the pseudo code of the DT classifier [33].

Algorithm 3. The pseudo code for the DT

*Input: Dataset C, R number of Instances, P features
 Output: Predicted class
 ConstructTree (R):
 if R contains instances of a single class then
 return
 else
 The feature P which has the greatest information gain is selected
 to split on
 Generate p leaf nodes of R ,
 where R has R_1, \dots, R_p and P has p possible values (P_1, \dots, P_p)
 for $i = 1$ to p do
 Define the content of R_i to C_i , where C_i is all the instances in R
 that match P_i
 Get ConstructTree (R_i)
 end for
 end if*

4. Support Vector Machine (SVM)

The SVM is a type of machine learning that information needed to determine into decision surfaces. It is used for both classification and regression problems. In this work, the SVM algorithm is used to categorize the result according to the input variables. The decision surfaces then break the data into two hyperplanar groups. [16]. The training data identify the vector that assists the hyperplane. Apparently, due to the larger margins, with a weak classifier generalization, a hyperplane that is farther away from the nearest training data point consistently has better margins and larger mistakes. Algorithm 4 gives the pseudo code of the SVM classifier [34].

Algorithm 4. The pseudo code for the SVM

*Input: Dataset C
 Output: Predicted Class
 Require: X and y uploaded with training labeled data, $\alpha \leftarrow 0$ or
 $\alpha \leftarrow$ partially trained SVM
 $A \leftarrow$ any value
 repeat
 for all $\{x_i, y_i\}, \{x_j, y_j\}$ do
 Optimize α_i and α_j
 end for
 until a change of α or other resource constraint criteria is not met
 Ensure: Remember only the support vectors ($\alpha_i > 0$)
 Test the model
 Calculate Scores
 Compute Confusion Matrix
 Validate Model*

5. Random Forest (RF)

The RF is a well-known and extensively used supervised machine learning approach to solve classification and regression issues [29]. The RF is an ensemble technique, and it combines several homogeneous learners as a single model. It uses decision tree algorithm for the prediction process, and it takes the final decision based on the average voting method. Algorithm 5 gives the pseudo code of the RF classifier [33].

Algorithm 5. The pseudo code for the RF

Input: Dataset C, R number of instances, P features
Output: Predicted class
 To create L classifiers
 for $i = 1$ to l do
 Randomly select the training data C with substitution to produce C_i
 Generate a parent node, R_i containing C_i
 Get Construct Tree (R_i)
 end for
 ConstructTree (R)
 if R contains instances of a single class then
 return
 else
 The $x\%$ of possible splitting features in R are randomly selected
 The feature P which has the greatest information gain is selected
 to split on
 Generate p leaf nodes of R, R_1, \dots, R_p ; where P has p possible values
 (P_1, \dots, P_p)
 for $i = 1$ to p do
 Define the content of R_i to C_i , where C_i is all the instances in R
 that match P_i
 end for
 Get ConstructTree (R_i)
 end for
 end if

6. Bagging

Bagging, also termed bootstrap aggregation, is a technique that was developed by Leo Breiman [26] to train and combine numerous homogenous learning algorithms. [13]. Bagging technique is used to reduce the problems related to overfit. Bagging is based on parallel method, and it uses data subsets for training the base learners. It optimizes the learning algorithm's robustness as well as the prediction algorithm's results [26]. It predicts the outcome with the help of voting method for classification. Since bagging does not allow recalculation of weight, changing the weight update equation is critical or reviews the algorithm's calculations. Algorithm 6 gives the pseudo code of the Bagging classifier [35].

Algorithm 6. The pseudo code for the Bagging

Input: T_1 : Training sample of C size dataset,
 s : count of bootstrap samples, L_c : Learning Classifier
Output: L^ bagging ensemble with s element classifiers*
 Learning stage:
 for $i = 1 \rightarrow s$ do
 $K_i \leftarrow$ bootstrap sample from C
 Create classifier $L_i \leftarrow L_c(K_i)$
 end for
 Predict the class label for a new sample
 $L^*(x) = \arg \arg \max_y \sum_{i=1}^s [L_i(x) = y]$

B. Characteristic Comparison of Each Classifier

This section discusses the pros and cons of each of the classifiers used for prediction. The kNN handles both classification and regression problems well but cannot deal with missing values. Though each feature makes unique assumptions about prediction outcomes, the NB is unaffected by irrelevant characteristics. While the DT provides feasible and adequate results for large data sources relatively rapidly, the algorithm must be trained over a long period of time and is also much more complex. Though the SVM is most effective at higher dimensions, it is vital to select a hyperparameter appropriately, and

there is no probabilistic explanation for the classification. The RF handles missing data very well, but overfitting occurs with noisy data. On huge datasets, the bagging approach performs well; nonetheless, there is a loss of interpretability in the model.

III. EXPERIMENTAL RESULT ANALYSIS & DISCUSSIONS

A. Dataset Description

This research utilizes two different types of datasets such as Crop and Prisoner's respectively. The details of these dataset are given in Table I.

TABLE I. DATASET DESCRIPTION

Dataset	Number of Instances	Number of Attributes	Type
Crop	1000	16	Nominal
Prisoner's	463	31	Numeric
Iris	150	4	Nominal

The crop dataset comprises soil and environmental factors, is downloaded from www.tnau.ac.in. The crop dataset has 1000 instances with 16 attributes in which 12 attributes are soil characteristics such as macro nutrients (nitrogen, phosphorus, potassium, etc.), macronutrients (zinc, iron, copper, etc.) and the remaining 4 are environmental such as rainfall, soil texture, temperature, and season. Also, this work utilizes to validate the performance of classifiers with other two dataset such as prisoner's and iris. Crime Propensity Prediction dataset [36] that can be used to predict the crime of a prisoner which was taken from the website github.com. The prisoner's dataset contains behavior of the prisoners with 463 instances and 31 attributes. Iris dataset [37] helps to find the iris plant class, which was downloaded from the University of California, Irvine. The dataset includes types of iris plant with 150 instances and 4 attributes.

B. Performance Metrics

The performance metrics namely, Accuracy, Kappa, Precision, Specificity, F1 Score, Area Under the Curve (AUC), and Mean Absolute Error (MAE) are used to predict the performances of each classifier. The formulae, and a representation of each metric used in the result examination, are stated in [38, 39].

C. Results and Discussion

In this section, the prediction performances of the classifiers are examined by various above mentioned performance metrics.

1. Sample Input and Output

Table II demonstrates the sample input and output range of the crop dataset, which includes the 12 soil characteristics of the potential of Hydrogen (pH), electrical conductivity (EC), organic carbon (OC), nitrogen (N), phosphorus (P), potassium (K), sulphur (S), zinc (Zn), boron (B), iron (Fe), manganese (Mn), and copper (Cu), as well as the 4 environmental characteristics of soil texture, seasons, rainfall, and average temperature. The expected output for the given input data, collected from the particular region, is given.

2. An Empirical Assessment of Classifiers Based on Soil Characteristics

Table III compares of the classifiers and identifies the best for appropriate crop. The process is based solely on soil characteristics like pH, N, and P.

The ensemble bagging classification technique clearly beats the others, as evidenced by the findings. Bagging also receives votes for increased performance for each sample, and as a result, it has a higher crop prediction accuracy than other approaches.

TABLE II. SAMPLE INPUT AND OUTPUT

Input																Output
pH	EC	OC	N	P	K	S	Zn	B	Fe	Mn	Cu	Texture	Season	Rainfall	Avg. Temp	
7.8	0.72	0.26	160	252.5	400	16	0.56	0.95	10.64	6.46	0.98	1	Kharif	296.8	25	Black Grams
7.8	0.72	0.26	160	252.5	400	16	0.56	0.95	10.64	6.46	0.98	1	Kharif	296.8	25	Black Grams
7.4	0.28	0.26	157.5	95.0	720	23	0.76	0.86	15.20	11.60	0.82	1	Rabi	296.8	25	Chick pea
7.9	0.73	0.31	175	267.5	400	31	0.54	0.84	9.86	6.36	1.02	1	Kharif	296.8	25	Maize
7.9	0.73	0.31	175	267.5	400	31	0.54	0.84	9.86	6.36	1.02	1	Kharif	296.8	25	Maize
7.4	0.28	0.26	157.5	95.0	720	23	0.76	0.86	15.20	11.60	0.82	1	Rabi	296.8	25	Chick pea
7.4	0.28	0.26	157.5	95.0	720	23	0.76	0.86	15.20	11.60	0.82	1	Rabi	296.8	25	Chick pea
8.2	0.14	0.20	140	97.5	972.5	13.3	0.78	0.49	9.50	5.54	0.74	1	Kharif	296.8	25	Maize
8.2	0.10	0.02	140	240	1000	2.34	0.82	0.34	10.40	5.54	1.08	1	Kharif	296.8	25	Maize

TABLE III. EMPIRICAL EVALUATIONS OF CLASSIFIERS BASED ON SOIL FACTORS

Classifiers	Performance Metrics							
	Accuracy (%)	Kappa (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)	AUC (%)	MAE
kNN	79.92	76.97	79.63	82.78	80.24	79.93	80.21	0.45
NB	80.65	78.07	81.03	83.90	81.25	81.13	81.06	0.37
DT	83.37	81.17	83.15	85.56	83.60	83.37	86.00	0.33
SVM	84.82	83.08	85.70	86.00	86.30	85.99	85.56	0.28
RF	88.82	87.35	88.73	90.27	89.79	89.25	89.19	0.20
Bagging	91.55	90.42	91.27	92.59	91.79	91.52	92.34	0.18

TABLE IV. EMPIRICAL EVALUATIONS OF CLASSIFIERS BASED ON ENVIRONMENT FACTORS

Classifiers	Performance Metrics							
	Accuracy (%)	Kappa (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)	AUC (%)	MAE
kNN	46.6	43.65	46.31	49.46	46.92	46.61	47.04	0.88
NB	47.33	44.75	47.71	50.58	47.93	47.81	48.12	0.78
DT	50.05	47.85	49.83	52.24	50.28	50.05	50.77	0.71
SVM	52.50	50.76	53.38	53.68	53.98	53.67	53.34	0.63
RF	54.50	53.03	54.41	55.95	55.47	54.93	55.00	0.57
Bagging	58.23	57.10	57.95	59.27	58.47	58.20	59.45	0.50

TABLE V. EMPIRICAL EVALUATIONS OF CLASSIFIERS BASED ON SOIL AND ENVIRONMENT FACTORS

Classifiers	Performance Metrics							
	Accuracy (%)	Kappa (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)	AUC (%)	MAE
kNN	77.73	75.78	78.44	81.59	79.05	78.74	79.43	0.40
NB	81.46	78.88	81.84	84.71	82.06	81.94	82.00	0.33
DT	84.18	81.98	83.96	86.37	84.41	84.18	85.07	0.27
SVM	86.63	84.89	87.51	87.81	88.11	87.80	87.83	0.20
RF	91.63	90.16	91.54	93.08	92.60	92.06	92.12	0.19
Bagging	93.36	92.23	93.08	92.32	93.12	93.10	94.89	0.12

3. An Empirical Assessment of Classifiers Based on Environment Conditions

Table IV depicts a comparison of the classification techniques that are exclusively based on environment factors like season, texture, average temperature and rainfall.

Table IV depicts the prediction rate based only on environmental characteristics. However, the bagging classification technique performs better than others, based only on environmental factors. In addition, bagging offers improved crop prediction accuracy because it uses multiple learning algorithms.

4. An Empirical Assessment of Classifiers Based on Soil and Environmental Characteristics

Table V shows a performance estimation of the classifiers, based on both soil and environmental factors, to find the right crop for cultivation in a specific area.

The information in Table V suggests that the prediction rate is higher with the combined features of both the soil and the environment, rather than that based solely on either of the two. Combining soil and environmental data, bagging provides more accurate prediction results than other classifiers. With the bagging technique's aggregation operation, the variance of estimation is greatly minimized.

TABLE VI. EMPIRICAL EVALUATIONS OF CLASSIFIERS FOR PRISONER'S DATASET

Classifiers	Performance Metrics							
	Accuracy (%)	Kappa (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)	AUC (%)	MAE
kNN	94.20	91.25	93.91	97.06	94.52	94.21	94.93	0.40
NB	94.93	92.35	94.59	97.46	94.81	94.69	95.60	0.30
DT	95.65	93.45	95.43	97.84	95.88	95.65	96.00	0.25
SVM	96.10	94.36	96.22	97.58	96.98	96.59	96.89	0.20
RF	97.10	95.63	97.01	98.55	97.18	97.09	98.07	0.13
Bagging	97.83	96.70	97.55	98.87	98.07	97.80	98.50	0.10

TABLE VII. EMPIRICAL EVALUATIONS OF CLASSIFIERS FOR IRIS DATASET

Classifiers	Performance Metrics							
	Accuracy (%)	Kappa (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)	AUC (%)	MAE
kNN	89.32	87.13	87.72	91.61	88.01	87.86	90.54	0.40
NB	91.43	89.00	91.00	94.55	92.11	91.55	92.77	0.31
DT	90.93	89.42	88.88	92.63	89.23	89.05	92.04	0.26
SVM	92.74	90.39	92.68	94.19	93.42	93.04	93.87	0.21
RF	94.32	92.42	93.12	96.39	94.93	94.01	96.49	0.15
Bagging	95.43	93.90	95.21	97.00	96.21	95.707	97.51	0.11

The results are evaluated with the real-world dataset discussed in section 3.1, using various classifiers. The results show that the ensemble technique provides the most accurate results of all.

5. Performance Evaluation of Classification Techniques for Prisoners Dataset

Further, the classifiers are tested, and their performance is verified with the prisoners' dataset downloaded from the GitHub website. Table VI presents the performance-wise results of the classification techniques, using the metrics discussed in section 3.3.

It is inferred that the ensemble learner gives better prediction accuracy, at 97.83%, than single learners. The bagging technique outperforms other classifiers in prediction. The ensemble technique combines two or more single prediction models for the best prediction rate. Since the performance of the bagging technique is unaffected by missing values, it works better than other techniques.

6. Performance Evaluation of Classification Techniques for Iris Dataset

Consequently, the performance of the classifiers is evaluated with the iris dataset which was taken from the UCI website to predict the class of iris plant. Table VII presents the performance-wise results of the classification techniques, using the metrics discussed in section 3.3.

It depicts that the ensemble learner gives better prediction accuracy, at 95.43%, than single learners. The bagging technique works well than other classifiers in iris plant prediction.

7. Empirical Evaluation of the Bagging Technique Using K-fold Validation

The results presented in Tables III-VII show that the bagging technique performs better than the rest. The best fold for the bagging technique is determined using the fold variation method. The fold method is used to evaluate the potential of each classifier for prediction process. The given dataset is divided into two subgroups, with the first ($k-1$) being used to train the classifier and the second (k^{th}) being used to examine the classifier.

Table VIII depicts a performance evaluation of the bagging classification technique for the crop, prisoners' and iris datasets, examining outcome prediction using several folds.

Table VIII presents the bagging technique performance for fold variation, with folds ranging from 10 to 90. The bagging technique performs better with 10 folds than any other. Performance is evaluated using the metrics given in section 3.3. The results show that the bagging technique achieved accuracy of 93%, 98% and 95% for the crop and prisoners' datasets, respectively.

8. Empirical Evaluation of the Bagging Technique Using Data Splitting Validation

To determine the best data splitting range, the bagging classifier's efficiency for the prediction process is assessed using the data splitting validation method. The graphical representation below displays a performance assessment of the bagging technique for the crop and prisoners' datasets, based on data fractionation, with ranges varying from 25% - 75% and 75% - 25%. Performance is assessed according to the metrics described in section 3.3.

Fig. 2 depicts the performance of the bagging classifier in forecasting an appropriate crop, using a crop dataset based on the data splitting validation method. From the results, it is observed that the 70% - 30% splitting range outperforms others.

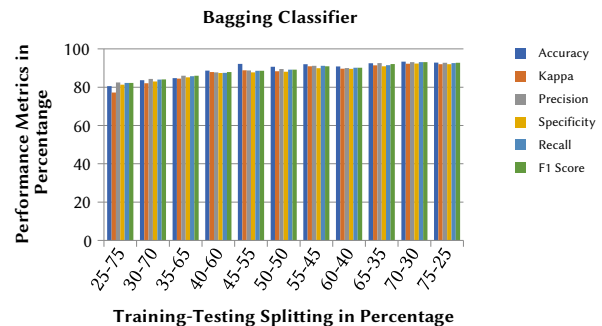


Fig. 2. Performance assessment of the Bagging classifier for Crop dataset utilizing data splitting method.

For the prisoners' prediction dataset, Fig. 3 illustrates a performance review of the bagging classification approach. The results show that the bagging technique outperforms the data splitting strategy in the 70% - 30% range.

TABLE VIII. PERFORMANCE OF THE BAGGING TECHNIQUE BASED ON FOLD VARIATION

Dataset	Folds	Performance Metrics					
		Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	Sensitivity (%)	F1 Score (%)
Crop	10	93.36	92.23	93.12	93.08	92.32	93.10
	20	92.04	90.91	91.80	91.76	91.00	91.78
	30	89.74	88.60	89.49	89.45	88.69	89.47
	40	91.06	89.62	90.51	90.47	89.71	90.49
	50	90.54	89.10	89.99	89.95	89.19	89.97
	60	89.24	87.80	88.69	88.65	87.89	88.67
	70	89.94	88.40	89.29	89.25	88.49	89.27
	80	90.51	88.97	89.86	89.82	89.06	89.84
	90	89.04	88.64	89.53	89.49	88.73	89.51
Prisoners	10	97.83	96.70	98.07	97.55	98.87	97.80
	20	96.51	95.38	96.75	96.23	97.55	96.48
	30	94.21	93.07	94.44	93.92	95.24	94.17
	40	95.53	94.09	95.46	94.94	96.26	95.19
	50	95.01	93.57	94.94	94.42	95.74	94.67
	60	93.71	92.27	93.64	93.12	94.44	93.37
	70	94.41	92.87	94.24	93.72	95.04	93.97
	80	94.98	93.44	94.81	94.29	95.61	94.54
	90	93.51	93.11	94.48	93.96	95.28	94.21
Iris	10	95.43	93.90	96.21	95.21	97.00	95.70
	20	94.87	93.56	95.65	94.97	95.78	95.30
	30	94.43	92.23	94.32	93.42	95.54	93.86
	40	95.11	93.45	95.48	94.51	96.18	94.99
	50	93.35	91.32	93.45	92.90	94.45	93.17
	60	93.66	92.45	94.12	92.39	94.43	93.24
	70	94.57	93.23	95.42	94.12	95.11	94.76
	80	92.14	91.45	93.04	92.04	93.34	92.53
	90	92.12	90.93	92.94	91.79	93.01	92.36

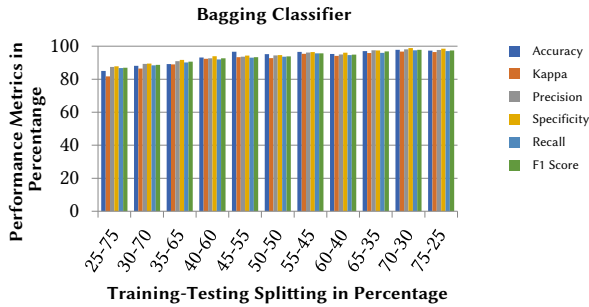


Fig. 3. Performance assessment of the Bagging classifier for Prisoners dataset utilizing data splitting method.

Fig. 4 shows a performance validation of the bagging classification method for the iris plant prediction dataset. The bagging technique performs better in the 70-30% data splitting range, according to the findings.

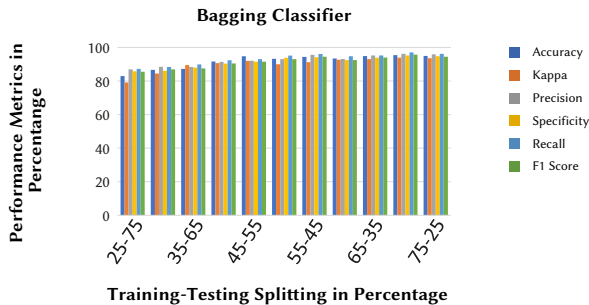


Fig. 4. Performance assessment of the Bagging classifier for Iris dataset utilizing data splitting method.

9. Empirical Evaluation of Each Classifiers Based on Time and Memory Occupation

Table IX illustrates the performance assessment of each classifier according to the execution time and memory of each.

TABLE IX. COMPARISON TABLE OF EACH CLASSIFIER BASED ON TIME AND MEMORY OCCUPATION

Classifiers	Time Taken (secs)	Space Occupied
kNN	0.19	260.72
NB	0.30	274.38
DT	0.47	261.45
SVM	0.29	292.87
RF	0.68	257.32
Bagging	0.70	246.01

It is evident from the results that though the bagging classifier requires a longer execution time than other techniques, it also occupies little space. It is inferred from Table 9 that the kNN classifier takes the lowest execution time with 260.72 KB of occupied space, while the SVM occupies the highest space but takes the second-lowest execution time.

IV. DISCUSSIONS AND CONCLUSION

A great deal of research has been carried out on the forecasting process using classification techniques. This work has examined the performance of the bagging, random forest, support vector machine, decision tree, Naïve Bayes and k-nearest neighbor classifiers using a crop dataset, a prisoners’ dataset and iris dataset. Using these algorithms, a relevant crop for cultivation was predicted from the crop dataset, the prisoners’ outcome predicted from the prisoners’ dataset, and the type of iris plant is predicted from the iris dataset.

The performance of the classifiers was examined using several performance metrics as accuracy, kappa, sensitivity, specificity, F1 score, area under the curve, precision, and mean absolute error. The results have shown that the bagging ensemble technique outperforms the rest. Then the bagging technique is examined by two validation methods namely fold and data splitting method. The obtained results show the bagging technique performs well on 10-fold and 70% - 30% data splitting range than others for predicting the target class of the given dataset.

REFERENCES

- [1] S. A. Z. Rahman, K. Chandra Mitra and S. M. Mohidul Islam, "Soil Classification Using Machine Learning Methods and Crop Suggestion Based on Soil Series," in *2018 21st International Conference of Computer and Information Technology (ICCIIT)*, Dhaka, Bangladesh, 2018, pp.1-4, doi: <https://doi.org/10.1109/ICCIIT.2018.8631943>.
- [2] William A. Belson, "Matching and prediction on the principle of biological classification," in *Journal of the Royal Statistical Society: Series C (Applied Statistics)* vol. 8, pp. 65-75, 1959, doi: <https://doi.org/10.2307/2985543>.
- [3] Richard O. Duda, Peter E. Hart, and David G. Stork. "Pattern classification and scene analysis", in New York: Wiley, vol. 3, 1973, doi: <https://doi.org/10.1086/620282>.
- [4] Gordon V. Kass, "An exploratory technique for investigating large quantities of categorical data", in *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 29, pp. 119-127, 1980, doi: <https://doi.org/10.2307/2986296>.
- [5] Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen, "Classification and regression trees", in *CRC press*, 1984, doi: <https://doi.org/10.1201/9781315139470>.
- [6] R.E. Neapolitan, "Models for reasoning under uncertainty", in *Appl. Artif. Intell.* vol. 1, pp. 337-366, 1987, doi: <https://doi.org/10.1080/08839518708927979>.
- [7] T. Kohonen, "Learning vector quantization", in *Neural Netw.* vol. 1, pp. 303, 1988, doi: https://doi.org/10.1007/978-3-642-97610-0_6.
- [8] Christopher G. Atkeson, Andrew W. Moore, and Stefan Schaal, "Locally weighted learning", in *Lazy learning*, Springer, Dordrecht, 1997, pp. 11-73, doi: <https://doi.org/10.1023/A:1006559212014>.
- [9] J Pearl, "Probabilistic Reasoning in Intelligent Systems", in *Morgan Kaufmann San Mateo*, vol. 88, pp. 552, 1988, doi: <https://doi.org/10.1016/C2009-0-27609-4>.
- [10] J.R. Quinlan, "C4.5: Programs for Machine Learning", in *Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA*, vol. 1, 1992, doi: <https://doi.org/10.1007/BF00993309>.
- [11] S.J. Russell, and P Norvig, "Artificial Intelligence: A Modern Approach", in Prentice Hall: Upper Saddle River, NJ, USA, vol. 9, 1995, doi: 10.1016/j.artint.2011.01.005.
- [12] L. Breiman, "Bagging Predictors", in *Mach. Learn.* vol. 24, pp. 123-140, 1996, doi: <https://doi.org/10.1007/BF00058655>.
- [13] Yoav Freund, and Robert E. Schapire, "Experiments with a new boosting algorithm", in *ICML 96*, pp. 148-156, 1996, doi: <https://dl.acm.org/doi/10.5555/3091696.3091715>.
- [14] Robert E. Schapire, "A brief introduction to boosting", in *IJCAI 99*, pp. 1401-1406, 1999, doi: <https://dl.acm.org/doi/10.5555/1624312.1624417>.
- [15] A. Smola, "Regression Estimation with Support Vector Learning Machines", in *Master's Thesis*, The Technical University of Munich, Munich, Germany, pp. 1-78, 1996.
- [16] Johan A.K. Suykens, and Joos Vandewalle, "Least squares support vector machine classifiers", in *Neural processing letters*, vol. 9, pp. 293-300, 1999, doi: <https://doi.org/10.1023/A:1018628609742>.
- [17] Leo Breiman, "Random forests", in *Machine learning*, vol. 45, pp. 5-32, 2001, doi: <https://doi.org/10.1023/A:1010933404324>.
- [18] J.A.K. Suykens, Van Gestel, T, De Brabanter, J, De Moor, B and J Vandewalle, "Least Squares Support Vector Machines", in *World Scientific: Singapore*, 2002, doi: <https://doi.org/10.1142/5089>.
- [19] Galvao, Roberto Kawakami Harrop, Mario Cesar Ugulino Araujo, Wallace Duarte Fragoso, Edvan Cirino Silva, Gledson Emidio Jose, Sofacles Figueredo Carreiro Soares, and Henrique Mohallem Paiva, "A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm", in *Chemometrics and intelligent laboratory systems*, vol. 92, pp. 83-91, 2008, doi: <https://doi.org/10.1016/j.chemolab.2007.12.004>.
- [20] M.S.P. Babu, N.V. Ramana Murthy and S.V.N.L. Narayana, "A web based tomato crop expert information system based on artificial intelligence and machine learning algorithms", in *International Journal of Computer Science and Information Technologies*, vol. 1, pp. 1-5, 2010, doi: 10.1.1.206.2072 .
- [21] S. Veenadhari, Bharat Misra, and C. D. Singh, "Machine learning approach for forecasting crop yield based on climatic parameters", in *2014 International Conference on Computer Communication and Informatics*, IEEE, pp. 1-5, 2014, doi: <https://doi.org/10.1109/ICCCI.2014.6921718>.
- [22] Paul Monali, Santosh K. Vishwakarma, and Ashok Verma, "Analysis of soil behaviour and prediction of crop yield using data mining approach", in *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, IEEE, pp. 766-771, 2015, doi: <https://doi.org/10.1109/CICN.2015.156>.
- [23] Jig Han Jeong, Jonathan P. Resop, Nathaniel D. Mueller, David H. Fleisher, Kyungdahm Yun, Ethan E. Butler, Dennis J. Timlin et al. "Random forests for global and regional crop yield predictions", in *PLoS One*, vol. 11, 2016, doi: <https://doi.org/10.1371/journal.pone.0156571>.
- [24] V. Sellam, and E. Poovammal, "Prediction of crop yield using regression analysis", in *Indian Journal of Science and Technology*, vol. 9, pp. 5, 2016, doi: 10.17485/ijst/2016/v9i38/91714.
- [25] S. Pudumalar, E. Ramanujam, R. Harine Rajashree, C. Kavya, T. Kiruthika, and J. Nisha, "Crop recommendation system for precision agriculture", in *2016 Eighth International Conference on Advanced Computing (ICoAC)*, IEEE, pp. 32-36, 2017, doi: <https://doi.org/10.1109/ICoAC.2017.7951740>.
- [26] Dipika H Zala, "Review on use of BAGGING technique in agriculture crop yield prediction", in *International Journal for Scientific Research & Development*, vol. 6, 2018.
- [27] Fabrizio Balducci, Donato Impedovo, and Giuseppe Pirlo, "Machine learning applications on agricultural datasets for smart farm enhancement", in *Machines*, vol. 6, pp. 38, 2018, doi: <https://doi.org/10.3390/machines6030038>.
- [28] Jahan Raunak, "Applying Naive Bayes Classification Technique for Classification of Improved Agricultural Land soils", in *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 6, pp. 189-193, 2018, doi: <https://10.22214/ijraset.2018.5030>.
- [29] P. Priya, U. Muthaiah, and M. Balamurugan, "Predicting yield of the crop using machine learning algorithm", in *International Journal of Engineering Sciences & Research Technology*, vol. 7, pp. 1-7, 2018, doi: 10.5281/zenodo.1212821.
- [30] A. Suresh, P. Ganesh Kumar, and M. Ramalatha, "Prediction of major crop yields of Tamilnadu using K-means and Modified KNN", in *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, IEEE, pp. 88-93, 2018, doi: 10.1109/CESYS.2018.8723956.
- [31] D. Anantha Reddy, Bhagyashri Dadore, and Aarti Watekar, "Crop Recommendation System to Maximize Crop Yield in Ramtek region using Machine Learning", in *International Journal of Scientific Research in Science and Technology*, vol. 6, pp. 485 - 489, 2019, doi: <https://doi.org/10.32628/IJSRST196172>.
- [32] Ivan Kholod, Andrey Shorov, and Sergei Gorlatch, "Improving Parallel Data Mining for Different Data Distributions in IoT Systems", in *International Symposium on Intelligent and Distributed Computing*, Springer, Cham, pp. 75-85, 2019, doi: https://doi.org/10.1007/978-3-030-32258-8_9.
- [33] Grant Anderson, "Random relational rules", in *PhD diss.*, The University of Waikato, 2008.
- [34] Angelina Tzacheva, Jaishree Ranganathan, and Sai Yesawy Mylavarapu, "Actionable Pattern Discovery for Tweet Emotions", in *International Conference on Applied Human Factors and Ergonomics*, Springer, Cham, pp. 46-57, 2019, doi: 10.1007/978-3-030-20454-9_5.
- [35] Mateusz Lango, and Jerzy Stefanowski, "Multi-class and feature selection extensions of roughly balanced bagging for imbalanced data", in *Journal of Intelligent Information Systems*, vol. 50, pp. 97-127, 2018, doi: <https://doi.org/10.1007/s10844-017-0446-7>.
- [36] H. Benjamin Fredrick David, A. Suruliandi, and S.P. Raja, "Preventing crimes ahead of time by predicting crime propensity in released prisoners using Data Mining techniques", in *International Journal*

of *Applied Decision Sciences*, vol. 12, pp. 307 – 336, 2019, doi: 10.1504/IJADS.2019.100433.

- [37] Dua, D. and Graff, C. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2019.
- [38] Mariammal, G., A. Suruliandi, S. P. Raja, and E. Poongothai. "Prediction of Land Suitability for Crop Cultivation Based on Soil and Environmental Characteristics Using Modified Recursive Feature Elimination Technique With Various Classifiers." in *IEEE Transactions on Computational Social Systems*, vol. 8, no. 5, 2021, pp. 1132-1142, doi: 10.1109/TCSS.2021.3074534.
- [39] Ganesan, Mariammal, Suruliandi Andavar, and Raja Soosaimarian Peter Raj. "Prediction of Land Suitability for Crop Cultivation Using Classification Techniques." in *Brazilian Archives of Biology and Technology*, vol. 64, 2021.
- [40] Bhaik, A., Singh, V., Gandotra, E., & Gupta, D. (InPress). Detection of Improperly Worn Face Masks using Deep Learning – A Preventive Measure Against the Spread of COVID-19. *International Journal of Interactive Multimedia and Artificial Intelligence*, In Press(In Press), 1-12. <http://doi.org/10.9781/ijimai.2021.09.003>
- [41] Alvarez, P., García de Quirós, J., & Baldassarri, S. (InPress). RIADA: A Machine-Learning Based Infrastructure for Recognising the Emotions of Spotify Songs. *International Journal of Interactive Multimedia and Artificial Intelligence*, In Press(In Press), 1-14. <http://doi.org/10.9781/ijimai.2022.04.002>
- [42] Sánchez-Torres, F., González, I., & Dobrescu, C. C. (2022). Machine Learning in Business Intelligence 4.0: Cost Control in a Destination Hotel. *International Journal of Interactive Multimedia and Artificial Intelligence*, 7 (Special Issue on Artificial Intelligence in Economics, Finance and Business), 86-95. <http://doi.org/10.9781/ijimai.2022.02.008>

G. Mariammal



G. Mariammal completed her B.E. degree in Computer Science and Engineering from Francis Xavier Engineering College, Tirunelveli, India, in 2011. She completed her M.E. degree in Computer Science and Engineering from Manonmaniam Sundaranar University, Tirunelveli, India, in 2017, also completed her Ph.D. degree in Computer Science and Engineering at Manonmaniam Sundaranar

University, Tamilnadu, India in 2021. Her research areas are machine learning, data analytics and image processing.

A. Suruliandi



A. Suruliandi completed his B.E. in Electronics & Communication Engineering in the year 1987 from Coimbatore Institute of Technology, Coimbatore. He completed his M.E. in Computer Science & Engineering in the year 2000 from Government College of Engineering, Tirunelveli. He obtained his Ph.D. in the year 2009 from Manonmaniam Sundaranar University, Tirunelveli. He is

working as a professor in the Department of Computer Science & Engineering in Manonmaniam Sundaranar University, Tirunelveli. He is having more than 29 years of teaching experience. He published 50 papers in International Journals, 23 in IEEE Xplore publications, 33 in National conferences and 13 in International conferences. His research areas are remote sensing, image processing and pattern recognition.

S. P. Raja



S. P. Raja was 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. His area of interest is image processing and cryptography. He is having more than 14 years of teaching experience in engineering colleges. Currently he is working as an Associate Professor in the school of Computer Science and Engineering in Vellore Institute of Technology,

Vellore. He published 42 papers in International Journals, 24 in International conferences and 12 in national conferences. He is an Associate Editor of the International Journal of Interactive Multimedia and Artificial Intelligence, Brazilian archives of Biology and Technology, Journal of Circuits, Systems and Computers, Computing and Informatics, International Journal of Image and Graphics and International Journal of Bio-metrics.



E. Poongothai

E. Poongothai completed her B.E. in 2011 from Anna University. She completed her M.E. and Ph.D., in computer science and engineering from Manonmaniam Sundaranar University, Tirunelveli, India, in 2013 and 2020 respectively. At present she is working as an Assistant Professor in the Department of Computer Science and Engineering, SRM University, Kattankulathur, Chennai.

Her research areas are Machine Learning and Computer Vision.

IoT Detection System for Mildew Disease in Roses Using Neural Networks and Image Analysis

Laura Torres¹, Luis Romero¹, Edgar Aguirre^{1*}, Roberto Ferro²

¹ Department of Informatic and Electronic, Faculty of Engineering, Minuto de Dios University Corporation - UNIMINUTO (Colombia)

² Francisco José de Caldas District University, Bogotá (Colombia)

Received 30 August 2021 | Accepted 9 May 2023 | Published 5 July 2023



ABSTRACT

Artificial intelligence presents different approaches, one of these is the use of neural network algorithms, a particular context is the farming sector and these algorithms support the detection of diseases in flowers, this work presents a system to detect downy mildew disease in roses through the analysis of images through neural networks and the correlation of environmental variables through an experiment in a controlled environment, for which an IoT platform was developed that integrated an artificial intelligence module. For the verification of the model, three different models of neural networks in a controlled greenhouse were experimentally compared and a proposed model was obtained for the training and validation sets of two categories of healthy roses and diseased roses with 89% training and 11% recovery. validation and it was determined that the relative humidity variable can influence the development and appearance of Downy Mildew disease when its value is above 85% for a prolonged period.

KEYWORDS

Classification, Convolution Neural Network, Images, Information System, Risk.

DOI: 10.9781/ijimai.2023.07.001

I. INTRODUCTION

NEW technological paradigms are impacting the agricultural area, such as artificial intelligence that is used to monitor relevant aspects of crops [1], generating new challenges towards the use of disease and pest detection systems, which are presented in correlation to other factors such as environmental ones, for which the use of different sources of information from different media allow enriching the analysis of certain problems that arise in the sector, two main elements are the analysis of images and the Internet of Things IoT, which they have become fundamental resources to understand the behavior of diseases and in the same way that of environmental variables, which, by correlating them, allow a better understanding of the appearance of diseases in crops.

In crops, different conditions can lead to the appearance of diseases and pathogens, especially in the crop of interest, which is roses. The appearance of this disease can affect productivity and in extreme cases can cause the total loss of a crop, due to the above, the need to detect the appearance of the downy mildew disease in time[2], is evident, to diagnose the state of a plant, helping the analysis and decision making through the analysis of images with neural networks, to rapid and accurate diagnosis of a culture.

Colombia is a Latin American country that is essentially agricultural where in recent years the gross domestic product of the agricultural sector has shown low growth, but during the 2012-2016 period it grew

on average 2.8%, compared to 4.2% of the national economy [3], and the increases that have occurred are due to the increase in production, the main flower exporters worldwide managed to produce US\$ 8,852 million, among which are the Netherlands (37%), Colombia (15.2%), and Ecuador (9.6%). Also, among the main importers worldwide are the USA (16%), Germany (15%), and the United Kingdom (13%) [4].

Likewise, in 2016 the main destination markets for flowers produced in Colombia were the United States with 76.2%, the United Kingdom 4.7%, and Japan 3.2%. According to the Colombian flower association, US\$1,312 million were exported. In this period, roses ranked first in exports with 20.5%, followed by carnations with 17.9%, chrysanthemums with 16.4%, alstroemeria with 8%, hydrangeas with 7.5%, and 29.8% from other species [4].

From the context of microclimatic variability, variables such as temperature and relative humidity that occur in greenhouses are related, therefore another challenge is to obtain those environmental variables that, if not known and controlled, can favor the appearance of downy mildew, especially in the rainy season and develops under certain environmental characteristics [5].

Downy mildew generates large economic losses year after year. Currently, the disease causes a 10% decrease in the total production of roses in the country [6], and losses reach up to 100% of flower stems [5]. Likewise, it significantly affects production if corrective measures are not taken in time, so it is necessary to anticipate the development of the disease due to its aggressiveness [5].

For this reason, in one way or another, flower growers carry out some type of monitoring, even if it is very simple, to detect the appearance of diseases and determine whether or not to apply phytosanitary treatments in time. However, they do not have the

* Corresponding author.

E-mail address: eaguirre@uniminuto.edu

appropriate methodology for the correct decision-making on several occasions since they are based on their knowledge and experience of the patients.

Now, the analysis of images and neural networks for the recognition of plant pests and diseases [7], proposes a new method for the detection of pests [8], and diseases [9], and the automatic supply of pesticides in greenhouse crops.

The detection of diseases in plants has been developed by different authors, in general, different types of classifiers are used to extract the particular characteristics, in the case of downy mildew in the context of corn [10] they have designed an image processing system of symptoms for disease detection using an ANN classifier.

Performance evaluation [11] is done through the K-nearest neighbor, naive bays (NB), LDA, and random forest tree (RFT) classifiers to classify diseases such as melanosis, greasy spot, and scabies. Likewise, the identification of foreign bodies [12] is used with a neural network in bulk food grains, having color and texture characteristics extracted from the sample images.

In the case of the evaluation of the quality of potatoes [13] an SVM classifier was implemented to detect defects, likewise, the same detection classifier [14] of disease spots is used.

In the case of coffee [15], it is used to quantify the severity of the rust disease by segmenting the infected areas from multispectral images, checking spots on the fruit, also in the processes [16] segmentation and feature extraction where it allows rapid and accurate detection of plant leaf diseases [17], the use of multispectral imaging [18] to classify leaf diseases is implemented particularly in cercospora beticola, and uromyces betae that is found in sugar beets.

In addition, another use of multispectral images [18] is for a K-NN classifier and a Bayesian classifier similar to [19] for the detection of defective apples due to scab, rot, and apple spot diseases, the above is based on k-means, which is also used in the detection and recognition of plant pests by k-means clustering and correspondence filter.

Due to the above, there is a problem in the early detection of the disease, taking into account that there are certain microclimatic conditions that allow it to reproduce, so the question of this work is in relation to how to detect the downy mildew disease. in roses? and the hypothesis is how to detect the disease using a neural network to analyze images and determine the occurrence of downy mildew disease in roses.

This research is focused on the development of a system for the detection of downy mildew disease in roses through image analysis using neural networks and the correlation of environmental variables through an experiment in a controlled environment, with the development of an IoT platform that integrates an artificial intelligence module, for the verification of the model, three different models of neural networks were compared with another to know the behavior of the system.

II. LITERARY REVIEW

A systematic analysis of the state of the art was carried out through five guiding questions that formed the research argument, the questions were: what was the objective of the investigations? what methodologies were used? in what context were the investigations carried out? investigations? what elements were used for the detection of diseases? And what kind of results did the authors obtain?

In the investigations found, different objectives were identified Fig. 1, which are described in this section, among which the advances in the different image processing techniques used for the study of plant diseases, traits, and pests are observed.

The recognition of diseases and pests was carried out through the implementation of recognition models [20] of diseases and pests, using the leaves of the plants based on images captured from different devices, and the automatic identification of plant diseases, based on visible range imaging [21].

The detection and classification [22], [23] of diseases are carried out through the automatic identification of diseases through image processing [24], [25]. The automatic identification [26] of the disease is also processed through artificial neural networks, from information crossed with a classifier [20] with PSO feature selection [26], it is also performed with disease detection (DDS) to detect and classify leaf diseases [27].

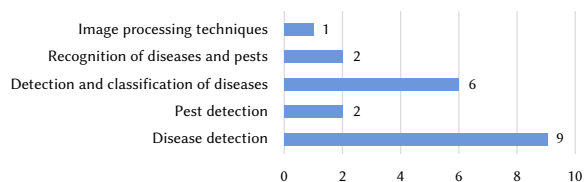


Fig. 1. Research objectives.

The methodologies and elements used are shown in Fig. 2, where one of the main elements is image acquisition [24]-[28], hence image preprocessing [29], [30], segmentation [29]-[31], comparison[32], [33], the analysis of histograms is performed and the use of detection techniques [26], [22] of edges that allow the training of the system [22], generate the knowledge base, perform the extraction of characteristics and classification, with Support Vector Machine, Matlab and neural networks.

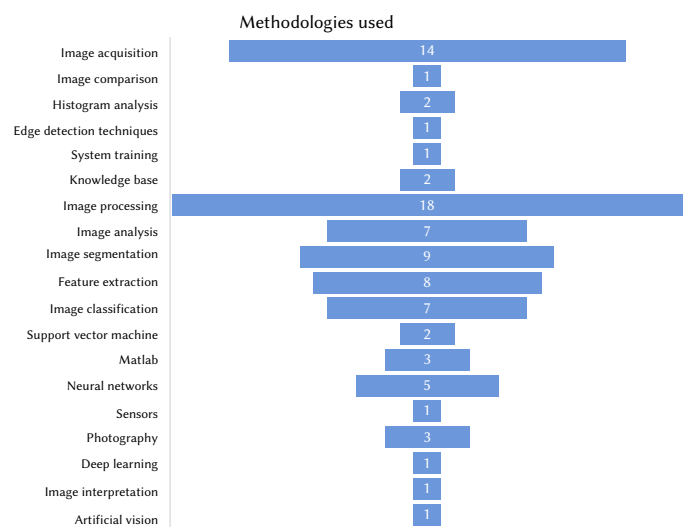


Fig. 2. Methodologies used.

The context in which the investigations are framed according to Fig. 3, is specifically divided into three parts, the first corresponds to the detection of diseases in different types of plants [21]- [34], where the rose is mentioned [35], [36]. The second is aimed at the detection of pests in different crops [37], [38]-[39], and finally, the improvement of yield in agricultural production is mentioned [40], [41]. Likewise, the investigations were carried out in different countries such as India, Mexico, Peru, and Ethiopia, and some of these in crops with controlled greenhouse environments [35] - [39].

The components and technologies used in the detection of pests and diseases Fig. 4, were mainly achieved with the implementation of image analysis, followed by image segmentation, neural networks,

and support vector machines (SVM). Without neglecting tools such as OpenCV, Image Bank, CANNY Borders, and Matlab.

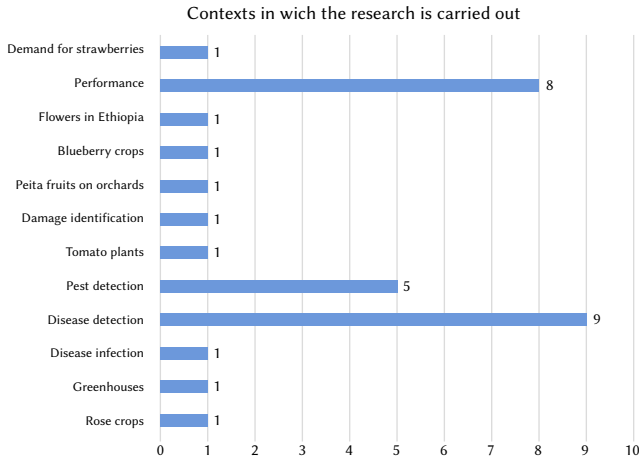


Fig. 3. Contexts in which the research is carried out.

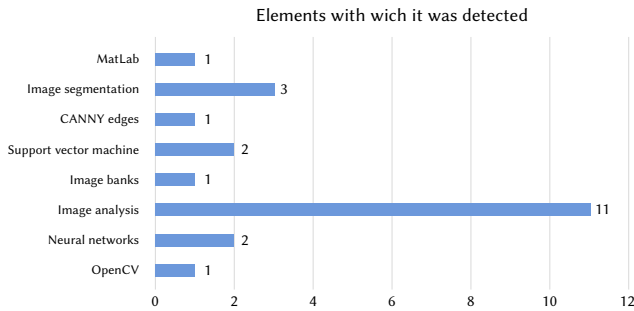


Fig. 4. Elements with which it was detected.

Of equal importance is to mention the devices and sensors that were used in the research work, where the common denominator was a simple digital camera, a cellular device, or the Everio JVC GZ-HD30 camera [35]. Additionally, a light sensor [35], was used to determine the intensity and obtain better quality images [2] as well as a scanner [39] allowing to improve the quality of the image for the detection of the target pest.

The results obtained by the investigations Fig. 5 are segmented into eight groups, within which the main one is oriented to the adequate detection of the disease [24], [42], [35], in second place is the precision in the disease identification [41], [36], [43], followed by: adequate detection of pests [20], [43], erroneous detection of diseases [28], images correctly classified with SVM [28], imágenes clasificadas correctamente con SVM [31], [32], [38], construction of hybrid systems between software and experts, reduction of human errors [37] and better performance of simple networks.

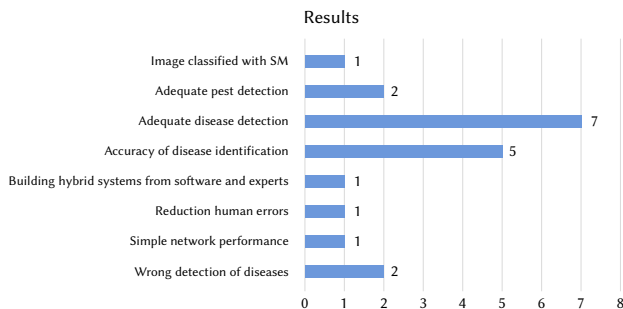


Fig. 5. Results obtained in the investigations.

III. PROPOSED METHOD

The developed system was framed in three components that allowed the experimental verification, Fig. 6, which are the on-site system, which corresponds to all the parts in charge of capturing the data of the microclimatic variables in the physical place to be monitored, the data processing, analysis, and administration is responsible for the storage, management of information and analysis through neural networks, the presentation is the last part and allows the user to access the information through a web interface.

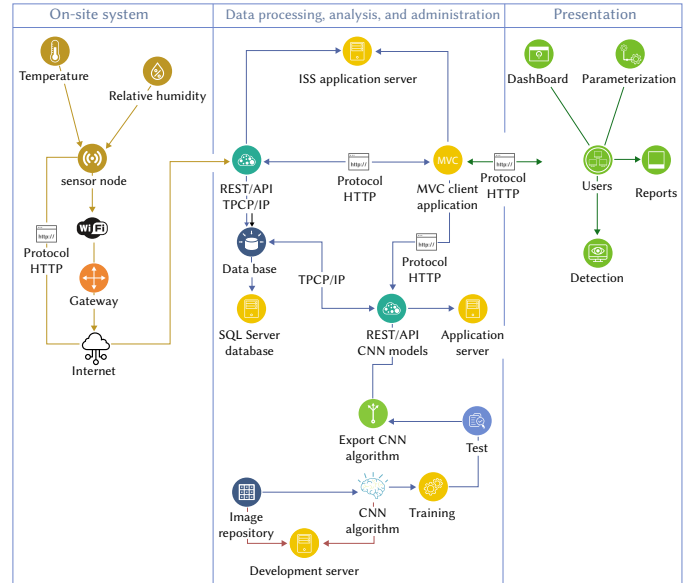


Fig. 6. Scheme of the system architecture.

The on-site system is made up of a microclimate station that integrates a temperature and relative humidity sensor, and has an internet connection, sending the captured data for storage is done through a REST/API, according to a period of time configured in the scheduled time. The installation of the station gives the user the possibility of obtaining data on the status of the variables automatically, to determine if at any time favorable conditions exist in which the rose bushes could be infected by the disease. downy mildew.

Data processing, analysis, and administration: it is the central component of the entire proposed solution since the network infrastructure and the application servers that support the operation of the system are located here; its parts are a database server in the SQL Server engine, an IIS application server that exposes all the REST/APIs that execute transactions with the database and additionally supports the client's web application, developed in ASP .Net MVC framework. On the other hand, in the detection of Downy Mildew disease based on the analysis of images or photographs, there is a machine that houses the development environment, where the training and testing of the convolutional neural network algorithms are carried out.

Likewise, an application server is integrated that exposes the REST/APIs with the Flask-Python application that gives access to the classification algorithms, stores the images in the database and returns the result of the classification thrown by each of the algorithms implemented towards the client application.

The presentation corresponds to the web interface that allows the user to view the information handled by the system and its configuration, the above, through its four main modules that are the DashBoard, where the data on changes in temperature and humidity relative to the on-site system captures them, and the alerts generated based on configured thresholds, the parameterization, here the user

makes the system configurations according to his personalization data taking into account the entity to which he belongs, monitoring devices which have, among others, the following module is the reports module that allows historical queries of the behavior data of the variables and alerts generated, the detection module enables a web interface with a form to upload an image in JPEG format, which is passed to the CNN algorithms and the answer is returned with two possible options, the first that the plant is sick or that is healthy.

IV. EXPERIMENT AND RESULTS

The data architecture was approached through the design of the component diagram Fig. 7, which contains five groups of components, starting first with the connection of monitoring devices, where there are three components corresponding to the on-site monitoring device, which communicates via WiFi to send data to the Internet, the second component is the business logic, where there is the REST APIs that allow transactions between the database and the presentation MVC application and finally the processing of Images where part of the example image repository that uses the classification model for training and validation, which is ultimately exported to be consumed from the REST API by the MVC application.

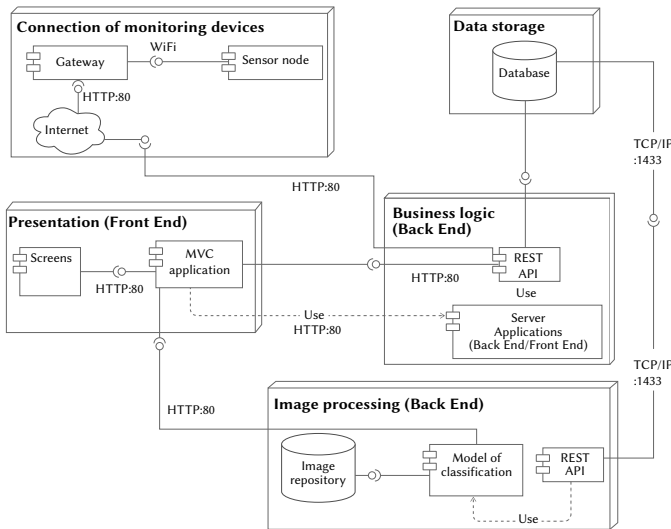


Fig. 7. Component diagram.

Second, the design of the Fig. 8 deployment diagram was made, which contains five groups of devices and a package corresponding to the IoT Hardware. The correct order for the implementation of the system begins with the implementation of the SQL Server, followed by the deployment of the IoT hardware, third the CNN model server is implemented, fourth the implements the back-end server, fifth the server is deployed to finally deploy the client.

Thirdly, the design of the entity-relationship model Fig. 9 was carried out, where eleven entities corresponding to the functionalities of the system with their respective attributes were established, the main entity is Image, which in turn is related to the Monitoring Device entity, which is related to the entities of Concentrator, User, Location and Variable, and the latter is related to Captured Data, and Threshold that is related to Alert and Disease.

Finally, the database was developed in the SQL SERVER database manager, considering the entity-relationship model, the database was composed of twelve tables, within which is the Image table, where each one was stored. of the images analyzed with the date and the result obtained by each of the four classification algorithms, in addition, the image with a Device Monitoring table that in turn is related to User,

Location, Concentrator, Threshold, VariableXDevice and Captured Data, these last two stores the variable assignment for each device and the data sent by the monitoring device respectively. In addition, they are related to the Variable table where the variables to be captured by the monitoring device are stored, each variable has many records in the Alert and Threshold table, which in turn are related to the Disease table.

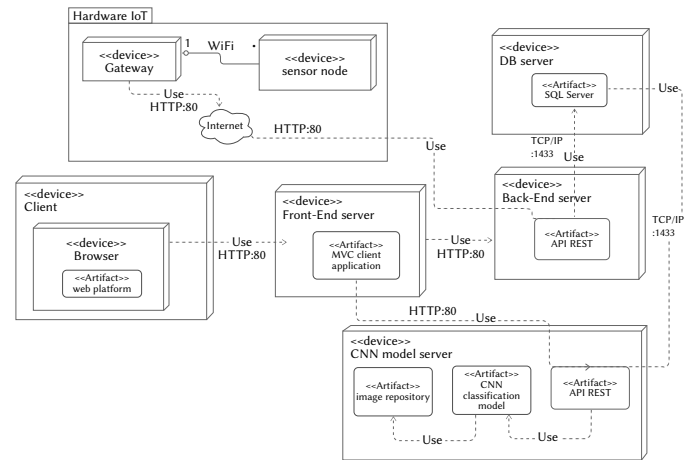


Fig. 8. Deployment diagram.

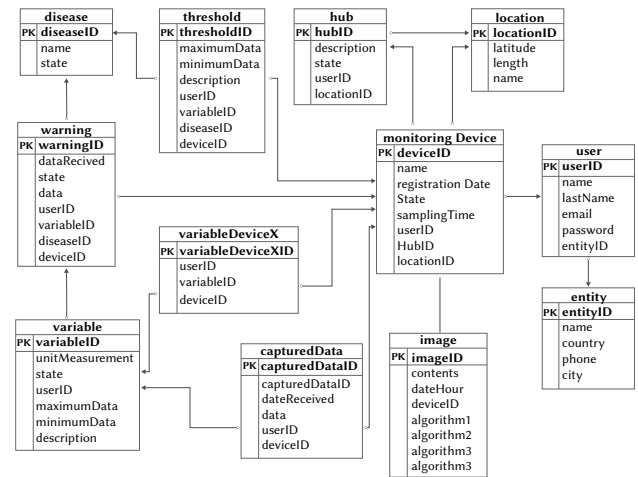


Fig. 9. Relational database diagram.

The development and coding of the IoT platform were carried out by implementing Microsoft technology in the ASP.Net framework with the C# programming language for the Front-End and Back-End, together with HTML, CSS, and JavaScript. Four modules were established based on the defined user stories.

The dashboard performs the real-time display of the data captured by the monitoring device, for the graphical representation meters from the Highcharts library were used, whose data was loaded from the JSON file returned by the REST API.

The parameterization module contains the general configurations of the system corresponding to the entity, company, or institution, location with geographic coordinates, hub, WiFi gateway or router, a monitoring device, microclimate variable, disease, and a maximum and minimum value that represents a characteristic. of a variable.

In the report module, there are the graphs corresponding to the reports generated by the system, among which are history, line graph, basic statistics, bar graph, heat map, behavior graph, alert history, table with the information of generated alerts, location of stations with a map implementing OpenStreetMap.

In the detection module, there is the upload section where the image is uploaded in .jpeg format, when you click query the image is passed to the REST API via an HTTP POST request as a file, then the REST API decodes and applies the pre-processing of the image, then passes it to the different algorithms and these return the sick or healthy response, which is encoded in JSON format so that the REST API returns it to the corresponding client.

The monitoring stations were two as in Fig. 10, both are the same in their construction, they receive the data from the sensors and transmit them to the internet using WiFi technology, each station used a pair of MCU esp32 nodes, based on the ESP32 LX6 microcontroller dual-core, 40 MHz and 520 KB SRAM with 802.11 bgN WiFi connectivity, HT40 transceiver and Bluetooth Low Energy (BLE), lithium battery charging circuit and CP2102 USB interface for connection to PC.

The function of the station was to capture the data from the sensors, the first is a temperature and relative humidity sensor, the Dht22 module was used, which has a relative humidity range between 0-99%, a temperature range between -40- 80°C, and a resolution of 8 bits, the second sensor was the TSL2561 which is a sensor that digitally measures the intensity of light ranging from 0.1 to 40,000 Lux, it operates between temperature ranges from -30 to 80 lux: from 0.1 to 40,000, operating voltage: from 2.7 to 3.6V, and manages an interface with I2C protocol, the use of this station and the transmission for data storage, which allows comparing the values of sensors against image analysis detection.

The station processes are described in algorithm I, where the “WiFi.h” and “SPI.h” libraries were imported, the access credentials to the platform were declared by the get post method and the network credentials SSID and password, the variables were defined as String of temperature, humidity and light, later the variables were verified and the communication through WiFi was initialized, the program verifies the transmission of the package, being the data previously stored temporarily in variables, which are declared in the variable fields of the HTTP hypertext transfer protocol, to build and send the message over the network.

Algorithm I. Algorithm monitoring system1 import libraries

Precondition:

- 1 **declare** network variables
- 2 **declare** temperature variable
- 3 **declare** humidity variable
- 4 **define** port temperature & humidity sensor
- 5 **define** type sensor
- 6 **create** an instance of temperature & humidity sensor
- 7 **Connect** to Wi-Fi network with SSID and password
- 8 **function:** connection
- 9 **set** serial connection
- 10 **call** begin to start sensor
- 11 **end function**
- 12 **function:** read sensors
- 13 **Read** humidity sensor
- 14 **Read** temperature as Celsius
- 15 **Check** if any reads failed and exit early
- 16 **Print** Failed to read from sensor
- 17 **return**
- 18 **send packets** POST HTTP
- 19 **end function**

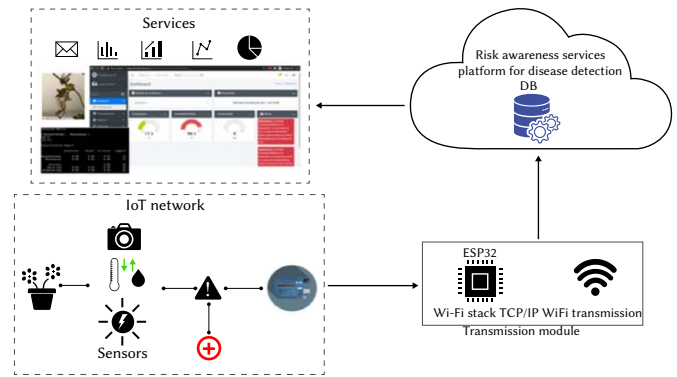


Fig. 10. Systems and subsystems of the hardware system implementation of the monitoring station.

In the context of the experiment, two scenarios were considered. The first consisted of a scale greenhouse Fig. 11a where a variety of healthy mini rose was planted, a humidifier, and a monitoring device that captured the microclimatic variables of temperature and relative humidity Fig. 11b. The second included a healthy mini rose variety in an outdoor environment and a monitoring device Fig. 11c.

The objective of the humidifier inside the greenhouse was to increase the percentage of relative humidity in the environment, to make the rose sick according to the values established in the state of the art of research for the appearance of Downy Mildew, and the monitoring device was included to be able to record the data in both scenarios.

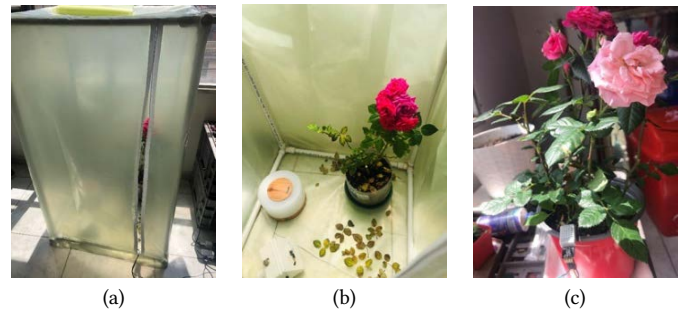


Fig. 11. (a) scale greenhouse (b) interior scale greenhouse (c) healthy rose exterior. (Author, 2021)

Within the experiment, photographs were taken at different times of the day for a week Fig. 12, both greenhouse rosebush and the outdoor rosebush, obtaining a total of 100 images with 50 sick and 50 healthy ones to upload them to the rose detection module system and also make the comparison concerning the temperature and relative humidity data captured by both monitoring devices.



Fig. 12. Sample of images taken.

The main element of objective three was to implement the digital image processing algorithm shown in Fig. 13.

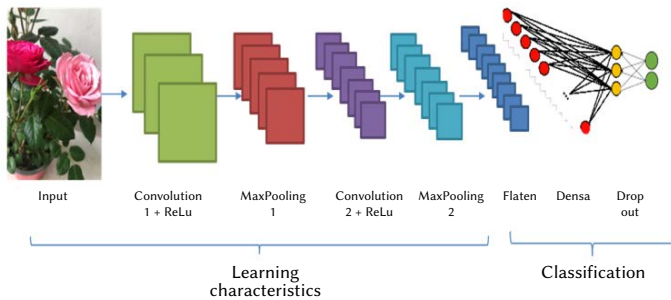


Fig. 13. Digital image processing algorithm.

Where the first thing that was carried out was the obtaining of images that allowed forming the set of images through algorithm II, for the training and testing of the four models to be implemented, a total of 200 images of different sizes were collected, one hundred are healthy rose bushes and one hundred are rose bushes that show visible symptoms of the disease with spots on the leaf.

Algorithm II. Digital image processing

Precondition:

- set epochs
- set height, length
- set batch size
- set steps
- set steps for validation
- set filters Conv
- set size filter
- set size pool
- set classes

```

1 function: Model
2   set cnn TO Sequential
3   add cnn ← filtersConv, size filter, activation
4   add cnn ← MaxPooling, size pool
5   add cnn ← Convolution, filters, size, activation
7   add cnn ← compile, loss, optimizers, accuracy
8 end function
9 function: Training
10  set fit cnn train, steps, epochs, validation
11 end function
    
```

The second is the pre-processing of the images, which mainly consisted of resizing them to a size of 32 x 32 pixels or 100 x 100 pixels in RGB format. The categories or classes where zero represents those that belong to diseased roses and one those that belong to healthy roses were also coded. In the third point, the Fig. 14 repository was segmented into images for training with 89% of the samples equivalent to 178 and the remaining 11% of the test equivalent to twenty-two. In this way, the feature matrix containing the pixels of each image that is passed to the algorithm for training and testing.

In the training process, the images are passed as an array of pixels through each of the layers of the CNN sequentially and in this way it learns the weights that allowed it to classify the new images. At the heart of the CNN are the convolution layers, which using multiple k

kernels allow to obtain significant characteristics of each image, such as the recognition of edges, points, objects, among others. An example is shown below in Fig. 15.

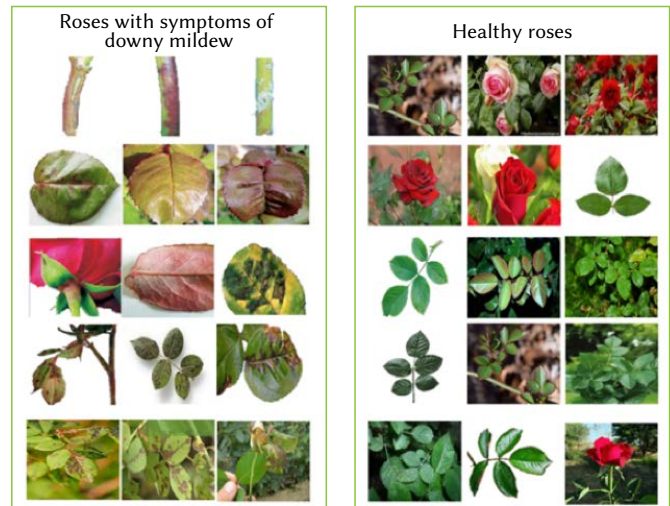


Fig. 14. Image repository sample.

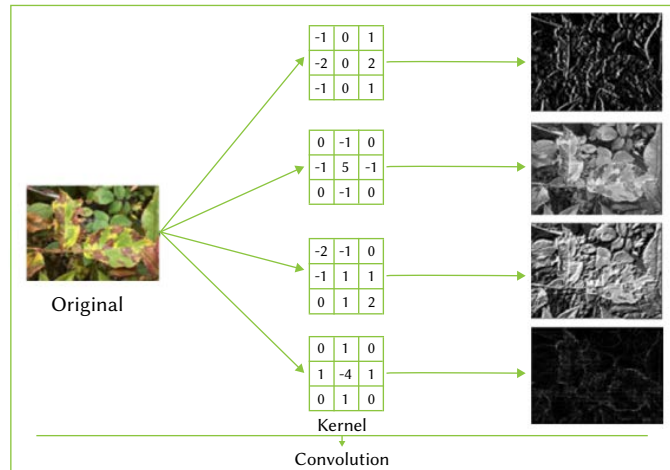


Fig. 15. Images after applying a convolution.

Finally, the results obtained from the training and testing processes of each model were evaluated to determine their quality in the classification of images representing a diseased plant. The metrics used are true positives (TP), which is the number of images that the model classified as rose plants with Downy Mildew and was correct, false positives (FP), which is the number of images that the model classified as rose plants with Downy Mildew. roses with Downy Mildew but that belong to images of healthy plants., true Negatives (VN) which is the number of images that the model classified as healthy rose plants and was correct, and false negatives (FN) which is the number of images that the model classified as healthy rose plants but belonging to images of rose plants with Downy Mildew.

Some of the metrics analyzed are found in Table 1 where precision, sensitivity and f1-score were obtained, where precision is the ratio between the number of true positives and the number of false positives. Precision is intuitively the classifier's ability to not label a negative sample as positive[44].

Sensitivity is the ratio of the number of true positives to the number of false negatives. Sensitivity is intuitively the classifier's ability to find all positive samples [44].

The F1-score can be interpreted as a weighted harmonic mean of accuracy and sensitivity, where an F-score reaches its best value of one and its worst score at zero [44].

TABLE I. CLASSIFICATION METRICS

Metric	Formula
Accuracy	$(\text{positives correctly classified} + \text{negatives correctly classified}) / (\text{positives correctly classified} + \text{false negatives} + \text{false positives} + \text{negatives correctly classified})$
Precision positives	$(\text{positives correctly classified}) / (\text{positives correctly classified} + \text{false positives})$
Precision negatives	$(\text{negatives correctly classified}) / (\text{positives correctly classified} + \text{negatives correctly classified})$
Sensitivity	$(\text{positives correctly classified}) / (\text{positives correctly classified} + \text{false positives})$
Specificity	$(\text{negatives correctly classified}) / (\text{positives correctly classified} + \text{false positives})$

The results obtained from the IoT platform are framed in the final modules, the reports generated from the data captured by each monitoring device, and the analysis produced by some of the uploaded images. Fig. 16 shows the real-time values of a controlled greenhouse.

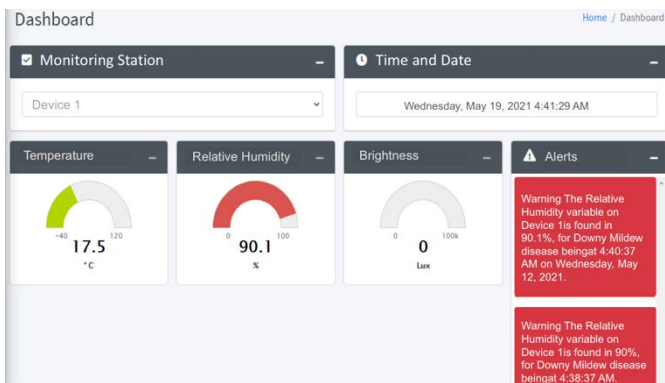


Fig. 16. Dashboard module.

From the environmental data obtained, an analysis is carried out through the stories to observe the behavior of the variables Fig. 17.

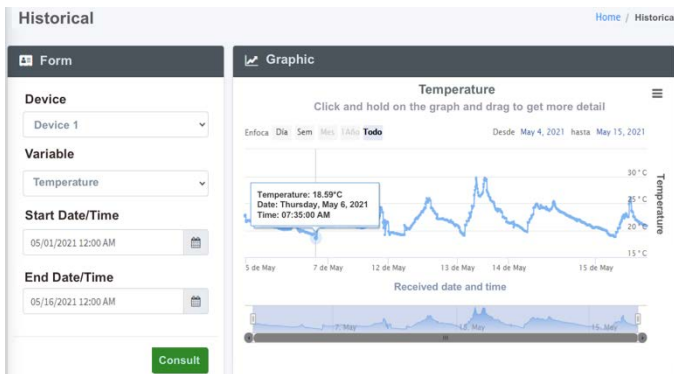


Fig. 17. History of environmental variables.

Likewise, the maximum, minimum, mode, mean, and mean values are determined in Fig. 18, which allows understanding the data set, in Fig. 19 a heat map is presented, and in Fig. 20 the loading of an image and the detection result.

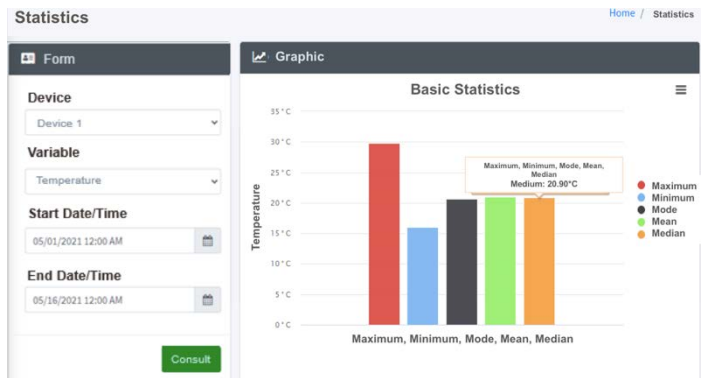


Fig. 18. Basic statistics.



Fig. 19. Heat map of the temperature variable.

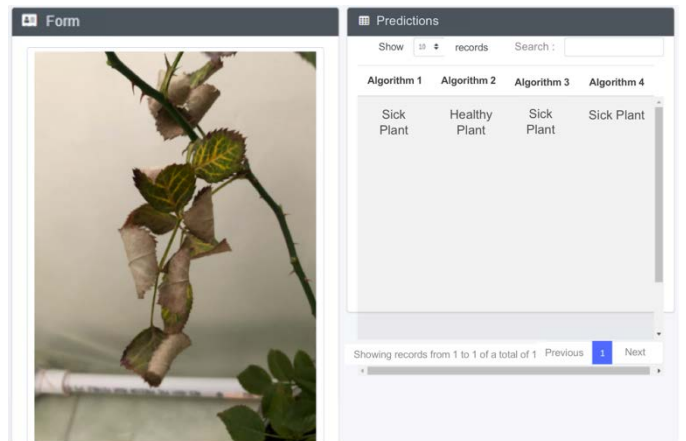


Fig. 20. Load image detection module.

To demonstrate the behavior of the microclimatic variables of temperature and relative humidity concerning the state of each rosebush, rose one corresponds to the one arranged in the greenhouse and rose two to the one located outside, to determine the effect of these on the state of the rose, a comparison was made of the data obtained by each monitoring device in a range of one week, which in this case are segmented as follows: device one corresponds to the sensor node located inside the greenhouse and device one device two corresponds to the sensor node located outside.

Fig. 21 represent the behavior and the temperature value that is repeated the most, respectively, for device one in a time range of one week. The most persistent data during the test was 22.50 °C, this being normal for the temperature required by the pink one.



Fig. 21. Device one temperature graph.

Fig. 22 represent the behavior and the most repeated humidity value, respectively, for device one in a time range of one week, the most persistent data during the test was 99.90% relative humidity. in the environment, this being very high and conducive to the appearance of Downy Mildew disease in the pink one.



Fig. 22. Humidity graph device one

Fig. 23 represent the behavior and the temperature value that is repeated the most, respectively, for device two in a time range of one week, the most persistent data during the test was 19.10°C, being this normal for the temperature required by pink two.



Fig. 23. Temperature behavior statistics

Fig. 24a shows the initial state of the rosebush before subjecting it to high relative humidity for a prolonged period of one week, and Fig. 24b shows the final state of the rosebush when observing the presence of Mildew Downy's disease.

Fig. 25a shows the initial state of rosebush two in an outdoor environment without alterations for one week and in Fig. 25b the final state of rosebush two is evidenced, observing that there are no changes in the state of the plant.



Fig. 24. (a) Start of pink test 1. (b) End of pink test 1.

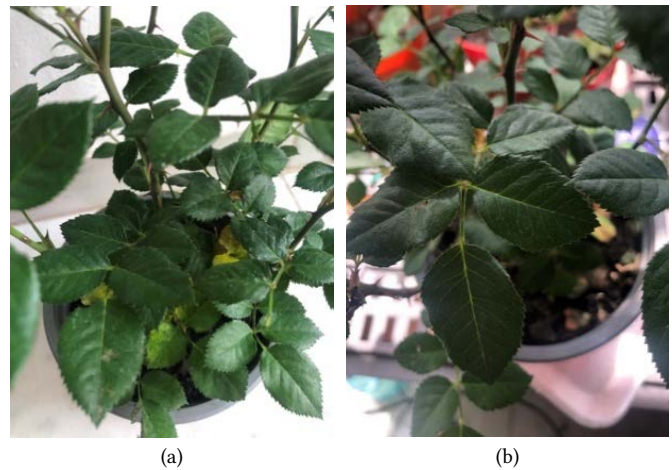


Fig. 25. (a). Start of pink test 2. (b) End of pink test 2.

V. ANALYSIS

About the data collected by the monitoring devices and the reports developed on the platform, the basic statistics graph was taken, which provides the mode information for each variable among other values, to observe what the value was of the data that was the most repeated during the execution time of the tests.

The mode value corresponding to the relative humidity for device one and rose one is outside the edaphoclimatic requirements of the rose, exposing it to an optimal level of humidity for the outbreak of Downy Mildew, which is triggered when the presence of the inoculum coincides. with high relative humidity above 85%, for a time greater than three hours according to [34] and when making the comparison between the beginning and the end of the state of the rose, it can be seen that indeed the alteration of humidity in levels as elevated for a long time, they can cause the appearance of Downy Mildew disease. Therefore, monitoring these microclimatic variables and generating alerts opens the possibility for the user to start monitoring with images for early detection of the disease.

The mode value corresponding to the relative humidity for device two and rosebush two is within the edaphoclimatic requirements of the rosebush, reducing the probability of the appearance of Downy Mildew disease as it is not exposed to microclimatic conditions that favor the appearance of the illness.

Regarding the results obtained in the different training and validation cycles, the CNN models were selected with the parameters of cycle four with input image size = $100 \times 100 \times 100 \times 3 = 30,000$, epochs = 100, and batch size = 32 since they were the ones that returned metrics closest to those expected, to obtain the greatest success in classifying new images in a real environment.

Table II presents the results classification report corresponding to the training stage of the final CNN models, to carry out a comparative analysis of them.

TABLE II. CLASSIFICATION REPORT

Shallow Net				
	Precision	Recall	F1-score	Support
Sick roses	0.67	0.55	0.6	11
Healthy roses	0.62	0.73	0.67	11
Accuarity	no data	no data	0.64	22
Macro avg	0.64	0.64	0.63	22
weighted	0.64	0.64	0.63	0.63
LeNet				
Sick roses	0.75	0.27	0.4	11
Healthy roses	0.56	0.91	0.69	11
Accuarity	no data	no data	0.59	22
Macro avg	0.65	0.59	0.54	22
weighted	0.65	0.59	0.54	22
MiniVGGNet				
Sick roses	0.75	0.55	0.63	11
Healthy roses	0.64	0.82	0.72	11
Accuarity	no data	no data	0.68	22
Macro avg	0.7	0.68	0.68	22
weighted	0.7	0.68	0.68	22
Proposal				
Sick roses	0.67	0.36	0.47	11
Healthy roses	0.56	0.82	0.67	11
Accuarity	no data	no data	0.59	22
Macro avg	0.61	0.59	0.57	22
weighted	0.61	0.59	0.57	22

In the first metric, training loss where the behavior is expected to decrease towards zero as times increase, which is an indicator that the model is learning from the example images, however, the results of all three. The former show pronounced variations in increase and decrease, since they show the described trend and reach a minimum value not less than 0.4.

This indicates that the learning process is carried out with errors, for the compensation of the error in the training set the images were repaired by the predictor and the ones that were detected were selected, to obtain the training set for which I review the result or output for each value \hat{y}_i and compare this value with the target value y_p , having as reference the subtraction $\hat{y}_i - y_p$, or the entire dataset, the total error committed was calculated as the sum of the individual errors. Therefore, the best images and weights were chosen to minimize the total error, which was initially 50%, but as the input data was improved and the algorithm was improved, this was substantially reduced. On the other hand, the proposed model shows a more stable downward trend, reaching a value of at least 0.01, which indicates a better learning rate.

The second metric is the validation loss, which is applied to the validation images and deviated the most from ideal behavior in all training cycles with each of the algorithms. The expected trend should decrease towards zero as the epochs increase, however, the result indicates a growth with drastic variations in increase and decrease indicating possible overfitting of the model, that is, that the model

only behaves well with the training images and when entering new ones it is not able to have an acceptable success rate in its classification as a sick rose or a healthy rose.

The third metric is the training accuracy where the behavior is expected to grow towards one as the epochs increase, which is an indicator that the model is learning from the example image features and classifying them correctly.

The fourth metric is the validation accuracy where the behavior is expected to grow towards one as the epochs increase, which would imply the correct classification of the images used in the validation set.

Table III presents the set of metrics of the confusion matrix and table IV the classification report of the results corresponding to the validation stage of the CNN models.

TABLE III. CONFUSION MATRIX TABLE

Model name	True Positive (TP)	False Positive (FP)	False Negative (FN)	True Negative (TN)
ShallowNet	6	5	3	8
LeNet	3	8	1	10
VGGNet	6	5	2	9
Proposed	4	7	2	9

TABLE IV. CLASSIFICATION REPORT CONFUSION MATRIX

Model name	Accuracy	Precision positives	Precision negatives	Sensitivity	Specificity
Shallow Net	0.6	0.5	0.7	0.6	0.6
LeNet	0.5	0.2	0.9	0.7	0.5
VGGNet	0.6	0.5	0.8	0.7	0.6
Proposed	0.5	0.3	0.8	0.6	0.5

Taking the confusion matrix as a reference, it was possible to show that the LeNet and proposed algorithms succeed and fail in the classification, obtaining a total of thirteen successes and nine failures out of twenty-two images belonging to the validation repository. On the other hand, the ShallowNet and MiniVGGNet algorithms achieve a total of fourteen and fifteen hits, respectively, so in the analysis of the confusion matrix the best algorithm is MiniVGGNet.

On the other hand, the classification report shows how the LeNet and proposed algorithms reach low insensitivity scores for the diseased roses category, indicating that the algorithms have a low ability to classify images belonging to this category. Regarding the ShallowNet and MiniVGGNet algorithms, it can be seen that the scores obtained for the metrics improve slightly to those of the LeNet algorithms and proposed highlights the MiniVGGNet algorithm.

The set of images that allowed the test to be carried out in a laboratory environment is made up of a total of one hundred, of which fifty are photos of the part where the plant presented the disease and fifty of the healthy part.

To begin with, based on the confusion matrix, it is possible to determine that the ShallowNet classifier is not capable of delivering the correct category for the images that correspond to diseased roses, since there is a total of true positives (TP) equal to zero and false positives (FP) equal to nine. Regarding the classification of healthy roses, he managed to hit forty-one (VN) and failed a total of fifty (FN). The LeNet classifier managed to reach a total of thirty-five (VP) diseased roses and a total of forty-eight failed (FP), a total of two (VN) healthy roses, and a total of fifteen failed (FN). Therefore, LeNet is a better classifier than ShallowNet.

Continuing, the MiniVGGNet classifier returned a total hit of fifty (VP) as diseased roses, however, it missed a total of forty-nine (FP), and in the category of healthy roses it managed to hit a total of one (VN), its

failure is equal to zero (FN). This provides an improvement over LeNet. Finally, the proposed classifier resulted in success in sick roses with a total of sixteen (VP) and failure in forty-one (FP), and in healthy roses, the classifier correctly returns nine (VN) but fails in thirty-four (FN). Thus, the proposed algorithm is below LeNet and MiniVGGNet in hits.

Finally, additional metrics such as the precision that represents the percentage of correct answers at a general level for all the samples confirm that, despite the failures in the MiniVGGNet classification, it is the one that performs best. Furthermore, it is reflected in the metrics not mentioned so far: the precision per category, the sensitivity (correct true positives), and the represented specificity (true negatives) confirm that MiniVGGNet was the most successful.

VI. CONCLUSION

The theoretical foundation on the different methodologies, components, and technologies implemented in the detection, classification of pests and diseases in crops, allowed a better perception of the operation, development, training, validation, and deployment of four models of convolutional neural networks, so addressing the solution of the problem from the creation of the set of images to obtaining the result of the experiment.

Including the development of a web platform in conjunction with a monitoring system for microclimatic variables that create environments conducive to the appearance of mildew adds value to the project since it allows alerts when the temperature and/or relative humidity are out of range. and initiate follow-up by taking images that are passed to classification models. To detect the disease in its initial stage, avoiding major effects on the cultivation of roses.

When carrying out the implementation of the CNN, the segmentation process of the image repository was evidenced, the extraction of characteristics of an image by executing an example with a 2D convolution layer that applied several k kernels that resulted in transformed images highlighting the characteristics of the edges and spots on the leaf of a rose. This allows CNN to learn the filters (kernel group k) to then classify the new input images.

As can be seen, in the results of the CNN in its training and validation stage, it was possible to show that all the graphs that show the loss and validation precision metrics have behaviors with drastic variations (increase and decrease) and the trend is opposite than expected, indicating possible overfitting in the learning performed by the models. Identified due to the few image samples that could be obtained for the training and validation sets (200 in total, in 2 categories: healthy roses and diseased roses with 89% training and 11% validation) and consequently, the training of the algorithms was not performed robust enough, making most results unacceptable.

According to the tests and the analysis of the results obtained with the microclimatic variables, it is observed that the relative humidity variable can influence the development and appearance of Downy Mildew disease when its value is above 85% during an extended period.

In the process of developing the research project, some future work fronts emerged related to neural networks and the implementation of an IoT platform for the storage of microclimatic data in real-time, since by merging these two modules it is possible to have a very robust system for the detection of any type of disease in plants. This is due to the detection module included in the CNNs and the possibility of real-time monitoring of the variables that affect a crop.

In the same way, the project can be improved by having a repository that has more examples of training and validation images and implementing other types of deeper CNNs that are not sequential, such as ResNet, ResNet50, among others, to compare them as well. with those implemented in this project.

Additionally, the monitoring device could be improved by including a camera as a sensor to take photos on-site and in real-time, which would allow a greater real knowledge of the current state of the crop.

REFERENCES

- [1] Y. Li, C. Xia, and J. Lee, "Vision-based pest detection and automatic spray of greenhouse plant," in *IEEE International Symposium on Industrial Electronics*, 2009, pp. 920–925. doi: 10.1109/ISIE.2009.5218251.
- [2] A. Calderón and H. Hurtado, "Vista de Machine learning en la detección de enfermedades en plantas," *Tecnología, investigación y academia TIA*, vol. 7, no. 2, pp. 55–62, Dec. 2019, Accessed: May 30, 2023. [Online]. Available: <https://revistas.udistrital.edu.co/index.php/tia/article/view/15685/15932>
- [3] F. Reyes, L. Cruz, N. Cáceres, and E. Valero, "Desempeño del sector floricultor," Bogotá D.C., 2017. Accessed: May 30, 2023. [Online]. Available: <https://sioc.minagricultura.gov.co/Flores/Normatividad/2016-06-01%20Boletin%20desempeño%20sector%20floricultor.pdf>
- [4] S. Johnny, "Ventaja comparativa del sector floricultor colombiano que promueva su presencia y le permita fortalecerse en el marco del TLC con corea del sur," Bogotá D.C., 2018. Accessed: May 30, 2023. [Online]. Available: <https://repositorio.uniagustiniana.edu.co/handle/123456789/370>
- [5] P. Israel *et al.*, "Current Status of Peronospora sparsa, Causal Agent of Downy Mildew on Rose (Rosa sp.) Estado Actual de Peronospora sparsa, Causante del Mildiu Velloso en Rosa (Rosa sp.)," *Rev. mex. fitopatol.*, vol. 31, no. 2, pp. 113–115, 2013, Accessed: May 30, 2023. [Online]. Available: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0185-33092013000200004&lng=es.
- [6] M. Ayala, I. A.; Luz, E. Argel-Roldan, S. Jaramillo-Villegas, M. Marín-Montoya, and M. M. Montoya, "Diversidad genética de peronospora sparsa (peronosporaceae) en cultivos de rosa de Colombia," *Acta biológica Colombiana*, vol. 13, no. 1, pp. 79–94, 2008, Accessed: May 30, 2023. [Online]. Available: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-548X2008000100005&lng=en.
- [7] A. Flórez, O. Hurtado, and S. Ramos, "Procesamiento de imágenes para reconocimiento de daños causados por plagas en el cultivo de Begonia semperflorens (flor de azúcar)," *Acta Agronomica*, vol. 64, no. 3, pp. 273–278, 2014, doi: 10.15446/acag.v64n3.42657.
- [8] C. Cáceres, D. Amaya, and O. Ramos, "Methodology for pest damage recognition in Begonia semperflorens link & Otto (sugar flower) crop through image processing," *Acta Agronomica*, vol. 64, no. 3, pp. 257–264, 2015, doi: 10.15446/acag.v64n3.42657.
- [9] F. Qin, D. Liu, B. Sun, L. Ruan, Z. Ma, and H. Wang, "Identification of Alfalfa Leaf Diseases Using Image Recognition Technology," *PLoS One*, vol. 11, no. 12, p. e0168274, Dec. 2016, doi: 10.1371/journal.pone.0168274.
- [10] S. A. Patil, D. S. Khot, O. D. Otari, and U. G. Malavkar, "Automated Disease Detection and Classification of Plants Using Image Processing Approaches: A Review," *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security*, vol. 203, pp. 641–651, 2021, doi: https://doi.org/10.1007/978-981-16-0733-2_45.
- [11] S. Reddy Bandi, "Performance evaluation of various statistical classifiers in detecting the diseased citrus leaves," *International Journal of Engineering Science and Technology (IJEST)*, vol. 5, no. 2, pp. 98–307, 2013, Accessed: Feb. 28, 2022. [Online]. Available: https://www.idc-online.com/technical_references/pdfs/information_technology/PERFORMANCE.pdf
- [12] U. Ansari, S. Shantaiya, and M. Ansari, "Identification Of Food Grains And Its Quality Using Pattern Classification," *International Journal of Computer and Communication Technology*, vol. 3, no. 1, 2012, doi: 10.47893/IJCCCT.2012.1107.
- [13] N. Razmjoooy, B. S. Mousavi, and F. Soleymani, "A real-time mathematical computer method for potato inspection using machine vision," *Computers & Mathematics with Applications*, vol. 63, no. 1, pp. 268–279, Jan. 2012, doi: 10.1016/J.CAMWA.2011.11.019.
- [14] S. Prasad, P. Kumar, R. Hazra, and A. Kumar, "Plant leaf disease detection using Gabor wavelet transform," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7677 LNCS, pp. 372–379, 2012, doi: 10.1007/978-3-642-35380-2_44.
- [15] D. Cui, Q. Zhang, M. Li, G. L. Hartman, and Y. Zhao, "Image processing

- methods for quantitatively detecting soybean rust from multispectral images,” *Biosystems Engineering*, vol. 107, no. 3, pp. 186–193, Nov. 2010, doi: 10.1016/J.BIOSYSTEMSENG.2010.06.004.
- [16] T. B. and S. V Rathod A N., “Image Processing Techniques for Detection of Leaf Disease,” *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 11, pp. 397–399, 2013.
- [17] S. D. Bauer, F. Korč, and W. Förstner, “The potential of automatic methods of classification to identify leaf diseases from multispectral images,” *Precision Agriculture*, vol. 12, no. 3, pp. 361–377, Jun. 2011, doi: 10.1007/S11119-011-9217-6/FIGURES/5.
- [18] A. Pacheco, H. Bolivar-Baron, R. Gonzalez-Crespo, and J. Pascual-Espada, “Reconstruction of High Resolution 3D Objects from Incomplete Images and 3D Information,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 2, no. 6, p. 7, 2014, doi: 10.9781/IJIMAI.2014.261.
- [19] S. R. Dubey, P. Dixit, N. Singh, and J. P. Gupta, “Infected Fruit Part Detection using K-Means Clustering Segmentation Technique,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 2, no. 2, p. 65, 2013, doi: 10.9781/IJIMAI.2013.229.
- [20] A. Devaraj, K. Rathan, S. Jaahnavi, and K. Indira, “Identification of plant disease using image processing technique,” *Proceedings of the 2019 IEEE International Conference on Communication and Signal Processing, ICCSP 2019*, pp. 749–753, 2019, doi: 10.1109/ICCSP.2019.8698056.
- [21] J. G. A. Barbedo, “A review on the main challenges in automatic plant disease identification based on visible range images,” *Biosystems Engineering*, vol. 144, pp. 52–60, 2016, doi: 10.1016/j.biosystemseng.2016.01.017.
- [22] M. Khari, A. K. Garg, R. Gonzalez-Crespo, and E. Verdú, “Gesture Recognition of RGB and RGB-D Static Images Using Convolutional Neural Networks,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 7, p. 22, 2019, doi: 10.9781/IJIMAI.2019.09.002.
- [23] Y. H. Robinson, S. Vimal, M. Khari, F. C. L. Hernández, and R. G. Crespo, “Tree-based convolutional neural networks for object classification in segmented satellite images,” *International Journal of High Performance Computing Applications*, Jul. 2020, doi: 10.1177/1094342020945026.
- [24] V. S. Bhong and P. B. V Pawar, “Study and Analysis of Cotton Leaf Disease Detection Using Image Processing,” *International Journal of Advanced Research in Engineering and Technology*, vol. 3, no. 2, pp. 1447–1454, 2016, doi: 10.1088/1742-6596/2062/1/012009.
- [25] R. Gupta, M. Khari, D. Gupta, and R. G. Crespo, “Fingerprint image enhancement and reconstruction using the orientation and phase reconstruction,” *Information Sciences (N Y)*, vol. 530, pp. 201–218, Aug. 2020, doi: 10.1016/J.INS.2020.01.031.
- [26] J. M. T. Ruiz, Jesús Gil and R. G. Crespo, “The Application of Artificial Intelligence in Project Management Research: A Review,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 54–66, 2021, doi: https://doi.org/10.9781/ijimai.2020.12.003.
- [27] K. Indumathi, R. Hemalatha, S. A. Nandhini, and S. Radha, “Intelligent plant disease detection system using wireless multimedia sensor networks,” *Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2017*, vol. 2018-Janua, pp. 1607–1611, 2018, doi: 10.1109/WiSPNET.2017.8300032.
- [28] G. Tigistu and Y. Assabie, “Automatic identification of flower diseases using artificial neural networks,” *IEEE AFRICON Conference*, vol. 2015-Novem, 2015, doi: 10.1109/AFRCON.2015.7332020.
- [29] V. Singh, Varsha, and A. K. Misra, “Detection of unhealthy region of plant leaves using image processing and genetic algorithm,” *Conference Proceeding - 2015 International Conference on Advances in Computer Engineering and Applications, ICACEA 2015*, pp. 1028–1032, 2015, doi: 10.1109/ICACEA.2015.7164858.
- [30] S. D. Khirade and A. B. Patil, “Plant disease detection using image processing,” *Proceedings - 1st International Conference on Computing, Communication, Control and Automation, ICCUBEA 2015*, vol. 7677, pp. 768–771, 2012, doi: 10.1109/ICCUBEA.2015.153.
- [31] J. N. Reddy, K. Vinod, and A. S. R. Ajai, “Analysis of Classification Algorithms for Plant Leaf Disease Detection,” *Proceedings of 2019 3rd IEEE International Conference on Electrical, Computer and Communication Technologies, ICECCT 2019*, pp. 1–6, 2019, doi: 10.1109/ICECCT.2019.8869090.
- [32] R. Meena Prakash, G. P. Saraswathy, G. Ramalakshmi, K. H. Mangaleswari, and T. Kaviya, “Detection of leaf diseases and classification using digital image processing,” *Proceedings of 2017 International Conference on Innovations in Information, Embedded and Communication Systems, ICIECS 2017*, vol. 2018-Janua, pp. 1–4, 2018, doi: 10.1109/ICIECS.2017.8275915.
- [33] F. Jobin, S. D. Anto, and B. K. Anoop, “Identification of leaf diseases in pepper plants using soft computing techniques,” *Conference on emerging devices and smart systems (ICEDSS)*, pp. 168–173, 2016, doi: 10.1109/ICEDSS.2016.7587787.
- [34] P. Revathi and M. Hemalatha, “Identification of cotton diseases based on cross information gain_deep forward neural network classifier with PSO feature selection,” *International Journal of Engineering and Technology*, vol. 5, no. 6, pp. 4637–4642, 2013.
- [35] N. Velázquez-López, Y. Sasaki, K. Nakano, J. M. Mejía-Muñoz, and E. R. Kriuchkova, “Detección de cenicilla en rosa usando procesamiento de imágenes por computadora,” *Revista Chapingo Serie Horticultura*, vol. 17, no. 2, pp. 151–160, 2011, Accessed: May 30, 2023. [Online]. Available: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1027-152X2011000200008&lng=es&tlng=es.
- [36] A. Fuentes, S. Yoon, S. C. Kim, and D. S. Park, “A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition,” *Sensors (Switzerland)*, vol. 17, no. 9, 2017, doi: 10.3390/s17092022.
- [37] S. Baskaran, “Advances in Image Processing for Detection of Plant Disease,” *The SIJ Transactions on Computer Science Engineering & its Applications (CSEA)*, vol. 05, no. 02, pp. 08–10, 2017, doi: 10.9756/sijcsea/v5i2/05010140101.
- [38] P. Rajan, B. Radhakrishnan, and L. Padma Suresh, “Detection and classification of pests from crop images using Support Vector Machine,” *Proceedings of IEEE International Conference on Emerging Technological Trends in Computing, Communications and Electrical Engineering, ICETT 2016*, 2017, doi: 10.1109/ICETT.2016.7873750.
- [39] P. Boissard, V. Martin, and S. Moisan, “A cognitive vision approach to early pest detection in greenhouse crops,” *Computers and Electronics in Agriculture*, vol. 62, no. 2, pp. 81–93, 2008, doi: 10.1016/j.compag.2007.11.009.
- [40] D. E. Kusumandari, M. Adzka, S. P. Gultom, M. Turnip, and A. Turnip, “Detection of Strawberry Plant Disease Based on Leaf Spot Using Color Segmentation,” *Journal of Physics: Conference Series*, vol. 1230, no. 1, 2019, doi: 10.1088/1742-6596/1230/1/012092.
- [41] R. G. De Luna, E. P. Dadios, and A. A. Bandala, “Automated Image Capturing System for Deep Learning-based Tomato Plant Leaf Disease Detection and Recognition,” *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, vol. 2018-October, no. October, pp. 1414–1419, 2019, doi: 10.1109/TENCON.2018.8650088.
- [42] L. Shanmugam, A. L. A. Adline, N. Aishwarya, and G. Krithika, “Disease detection in crops using remote sensing images,” *Proceedings - 2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development, TIAR 2017*, vol. 2018-Janua, no. Tiar, pp. 112–115, 2018, doi: 10.1109/TIAR.2017.8273696.
- [43] C. Sulca, C. Molina, C. Rodríguez, and T. Fernández, “Detección de enfermedades y plagas en las hojas de arándanos utilizando técnicas de visión artificial,” *Perspectivas*, vol. 15, no. 15, pp. 32–39, 2018.
- [44] A. Navlani, “Naive Bayes Classification using Scikit-learn,” *Datacamp*, 2020. <https://www.datacamp.com/tutorial/naive-bayes-scikit-learn> (accessed May 30, 2023).



Laura Estefania Torres Tovar

Systems engineer from the Corporación Universitaria Minuto de Dios-UNIMINUTO, works as a full-stack developer for private companies.



Luis Carlos Romero Cardenas

Systems engineer from the Corporación Universitaria Minuto de Dios-UNIMINUTO, works as a full-stack developer for private companies.



Roberto Ferro Escobar

Roberto Ferro Escobar is an Electronics Engineer from the Francisco José de Caldas District University, a master's in information and communication Sciences, and a Ph.D. in Computer Engineering from the Pontifical University of Salamanca. Full Professor at the same University. Researcher in HPC, Data Science, and Bioinformatics.



Edgar Alirio Aguirre Buenaventura

Edgar Aguirre He is a PhD in Engineering from the Francisco José de Caldas District University, master's in information and communication Sciences, Electronic Control Engineer, Electronics Technologist, works as director of the INDEC community development engineering institute, and is Professor in the Electronic Technology Program of the Corporación Universitaria Minuto de Dios-UNIMINUTO.

Quantitative Measures for Medical Fundus and Mammography Images Enhancement

Monserrate Intriago-Pazmiño^{1,2*}, Julio Ibarra-Fiallo³, Adán Guzmán-Castillo², Raúl Alonso-Calvo¹, José Crespo¹

¹ Biomedical Informatics Group, Universidad Politécnica de Madrid, Madrid (Spain)

² Departamento de Informática y Ciencias de la Computación, Escuela Politécnica Nacional, Quito (Ecuador)

³ Colegio de Ciencias e Ingeniería, Universidad San Francisco de Quito, Cumbayá (Ecuador)

Received 25 June 2021 | Accepted 23 November 2022 | Published 19 December 2022



ABSTRACT

Enhancing the visibility of medical images is part of the initial or preprocessing phase within a computer vision system. This image preparation is essential for subsequent system tasks such as segmentation or classification. Therefore, quantitative validation of medical image preprocessing is crucial. In this work, four metrics are studied: Contrast Improvement Index (CII), Enhancement Measurement Estimation (EME), Entropy EME (EMEE), and Entropy. The objective is to find the best parameters for each metric. The study is performed on five medical image datasets, three retinal fundus sets (DRIVE, ROPFI, HRF-POORQ), and two mammography image sets (MIAS, DDSM). Metrics are calculated using a binary mask image to discard the background. Using the fundus and mask datasets, the best results were obtained with the EMEE and EME metrics, which achieved mean improvements of up to 186% and 75%, respectively. For mammography datasets and using masks of the region of interest, the two metrics with the highest percentage improvement were CII and EMEE, which obtained means of up to 396% and 129%, respectively. Based on the experimental results provided, we can conclude that EMEE, EME, and CII metrics can achieve better enhancement assessment in this type of medical imaging.

KEYWORDS

Contrast Improvement Metrics, Contrast Quantitative Measures, Fundus Images, Mammography Images, Medical Image Enhancement.

DOI: 10.9781/ijimai.2022.12.002

I. INTRODUCTION

MEDICAL imaging is of great importance in helping specialists to diagnose many diseases. In principle, the specialist analyzes the image, sometimes with the aid of a computer-assisted diagnosis (CAD) system. These CAD systems belong to the field of computer or artificial vision. Artificial vision systems (AVS) for image analysis usually have the phases of preprocessing, segmentation, postprocessing, and feature computation or defect classification [1]. The preprocessing step produces an image of better quality or in a suitable condition for computational analysis in the following phases. The research of this paper focuses on the preprocessing phase. The objective is to study quantitative metrics that measure how much a preprocessed medical image has been improved. The behavior of four metrics on two types of medical images is studied.

The images are enhanced through several filters that modify their contrast and brightness to distinguish the parts of interest. Usually, a qualitative visual analysis is carried out in the preprocessing step, while the validation is focused on later phases. If the outcome is not as expected, the researchers may be forced to go back to the initial phase and try other preprocessing filters.

Concerning the impact of preprocessing tasks on a computer vision system performance, the Master's Thesis [2] examined the effect of preprocessing algorithms on the performance of Convolutional Neural Networks (CNNs) including transfer learning to detecting pneumonia and classifying cats or dogs. This research was conducted using the original images and five enhancement algorithms: the contrast limited adaptive histogram equalization (CLAHE), the successive means of the quantization transform (SMQT), the adaptive gamma correction, the wavelet transform, and the Laplace operator. The chest X-ray and pets' datasets were acquired from Kaggle Challenge. The results reported that LeNet5 CNN performance was improved with some enhancement algorithms, but transfer learning performance slightly decreased with pre-trained VGG16 CNN for pet images.

In contrast, the work [3] obtained better performance when its transfer learning model was tested with enhanced images. The authors studied the impact of enhancing chest X-ray images for COVID-19 detection by applying transfer learning. This study used the large X-ray dataset COVQU, which contains 18,479 CXR images where 8,851 are normal, 6,012 are non-COVID lungs, and 3,616 are COVID-19 CXR images. Additionally, COVQU includes the ground truth lung masks. The study involves classifying each image as normal, lung opacity, or COVID-19 with and without image enhancement. Five image enhancement procedures were used: histogram equalization (HE), contrast limited adaptive histogram equalization (CLAHE), image complement, gamma correction, and balance contrast enhancement technique (BCET). Transfer learning was carried out from six pretrained Convolutional

* Corresponding author.

E-mail address: monserrate.intriago@epn.edu.ec

Neural Networks (CNNs): ResNet18, ResNet50, ResNet101, InceptionV3, DenseNet201, and ChexNet, and the last eleven layers were re-trained. This study achieved seventy-two experiment settings (six pre-retrained CNNs with two datasets, and each dataset has been tested with no enhancement and with five different enhancing methods). The result showed that the image enhancement preprocessing improved the classification performance. Original images (without enhancement) were classified using InceptionV3 obtained 93.46% on accuracy average, and the best experiment setting using gamma correction and ChexNet reached 96.29% on accuracy average.

Other studies also reported higher performance after including some enhancement tasks. A pedestrian detection model based on the YOLOv3 convolutional neural network architecture that analyzes outcomes with and without preprocessing phase is presented in [4]. The preprocessing contrast enhancement is achieved using the Retinex method. The entire proposed model obtained 90% and 94% accuracy, without and with preprocessing, respectively.

In [5], a convolutional neural network (CNN) proposal for recognizing six basic emotions is implemented using several preprocessing methods. The main intention of this study is to investigate how preprocessing practices affect CNN performance. Face detection using a single pre-processing phase achieved a significant result with 86.08 % accuracy. However, combining some techniques increased the performance of CNN and achieved 97.06% accuracy.

The importance of the preprocessing phase in a computer vision system for detecting melanoma has been studied in [6]. Authors recall that preprocessing is the first and fundamental step for improving image quality. In this AVS, preprocessing is designed for removing noise and irrelevant portions against the background of skin photographs. This study applied different pre-processing methods utilized on skin cancer photos. These studies experimentally validated that a suitable preprocessing could increase accuracy.

Most of the research mentioned so far has experimented with the effect of enhancing the images before using them in the learning model of the computer vision system. On the other hand, the image enhancement was not quantified, or at least, it is not reported. To overcome this non-objective scenario, a quantitative image evaluation becomes significant. This research aims to analyze the behavior of quantitative metrics. As a result of this study, this paper recommends appropriate metrics for evaluating the enhancement in each type of image, and studies the values of the relevant algorithms' parameters.

The scientific literature is reviewed in this work, and several metrics used to quantify contrast in medical imaging are studied. Four metrics have been chosen: Enhancement Measure Estimation (EME), Enhancement Measure Estimation by Entropy (EMEE), Contrast Improvement Index (CII), and Entropy. The metrics are applied to two cases of studies: fundus and mammography images, on five data sets. These datasets contain healthy and pathological images. Another point is that some of these datasets include images of poor quality.

The remainder of this article is organized as follows. Section 2 presents a review of related works. Section 3 describes the metrics and their theoretical foundations, the types of images used, and the preprocessing algorithms. Section 4 shows the analysis of the parameters of each metric and its behavior on each dataset. Section 5 provides a discussion of the obtained results. Finally, some conclusions are given.

II. BACKGROUND

This section presents a review of research works that use metrics to evaluate the performance of the preprocessing filters. Works are presented according to the type of images. First, a cellular medical image work is analyzed; after that, some results corresponding to

mammograms are described; finally, studies of retina images of prematurely born children are included.

The work in [7] enhances the contrast of several types of cellular medical images. The enhancement performance is quantified using the CII measure. Cellular images can present complex shapes and textures. Thus, the research concludes that CII is a suitable measure for analyzing the enhancement of these types of images.

A recoloring algorithm, named RGBeat, is presented in [8] in order to assist patients with protanopia and deuteranopia. It is applied to images and text. The proposal uses CIE and RGB color spaces. This method converts a range of values of the hue channel to achieve a better understanding by these types of patients, while preserving the main characteristics. A validation of the modified image is performed. Consistency is addressed by ensuring that all pixels of the same color in an input image will have the same output color after applying the recoloring method. In addition, naturalness is measured by using a quadratic difference in the CIE Lab color space [9]. A small value indicates that the naturalness has been maintained in the recolored image. And the altered contrast measurement is based on a squared Laplacian [10].

The research in [11], which applies deep neural networks for diagnosing congenital heart diseases (CHDs) using echocardiographic ultrasound images, considers the possibility of preprocessing the ultrasound contrast. The authors mention that they are motivated by the work presented in [3], that demonstrated the improved performance of learning methods with contrast-enhanced images.

The contrast of mammography images is enhanced and validated using the CII measure [12]. This research presents a method based on the following filters: Laplacian Gaussian, Contrast Limited Adaptive Histogram Equalization (CLAHE), and morphological filters (openings and closings) to improve the contrast.

A metric based on the high and low-frequency range is proposed in [13]. This metric is used for assessing contrast quality in mammographies. According to the experiments, when the original image is compared with its enhancement, specific conditions of the frequency values are met. The contrast enhanced image has better quality. A private mammographies dataset was used. It contained 179 images, among benign, malignant, and normal mammographs. The study concluded that using multiple filters produces better results and that filter behavior varies between image types.

The machine vision system proposed in [14] aims to recognize the area of distortions in breast images. The distortion classification is mainly performed through an improved pulse-coupled neural network (PCNN). This research utilized the Digital Database for Screening Mammography (DDSM) dataset. In the preprocessing phase, the filters used are top-bottom hat and gamma transformation. Their improvement is validated through the Equivalent Number of Looks (ENL) and Contrast Improvement Index (CII) metrics.

Some measures in [15] are studied to validate contrast enhancement of mammography and tomography pathological images. In this study, ten images for each case are taken. Results are shown in terms of the metrics: EME, EMEE, Logarithmic Michelson Contrast Measure (AME), Logarithmic AME by Entropy (AMEE), Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Absolute Mean Brightness Error (AMBE), Discrete Gray Level Energy (GLE), Relative Entropy (RE), Second-Order Entropy (SOE), Edge Content (EC). The authors classified EME and EMEE as Complex Measures of Contrast, which are convenient metrics for medical images where the background is uniform.

In [16], a method for adjusting the contrast level in a windowed mammogram is proposed. The technique is called GRAIL (Gabor-Relying Adjustment of Image Levels), and it is regulated by a measure of mutual information (MI). MI is the relation between the decomposition

of the original 12-bit inputs and its screen-displayed 8-bit version. The mutual information metric between the original instance and its Gabor-filtered derivations is applied to X-ray images [17], and the results show a better performance in terms of subjective interpretation.

Regarding retinal studies, a new CLAHE version method is presented in [18], and its improvement is evaluated using the Entropy measure, tested on the public DRIVE and STARE fundus image datasets. Those datasets contain adult fundus images of normal and pathological cases.

A private fundus images dataset is used in the research reported in [19]. Authors apply a guided filter for the preprocessing phase of fundus images of children. Healthy and pathological images are used. This filter is assessed through EME, Entropy, Standard Deviation, and Spatial Frequency.

A second work that uses a private set of children's fundus images is [20]. Contrast is enhanced using adaptive filters based on the features of each image. The parameters for each filter are computed employing an artificial neural network. The enhancement is validated using CII and qualitative analysis.

Based on the related works that could be identified at the time of performing this work, the following section details the metrics that are tested, and the datasets of the two case studies that are the object of this research.

III. MATERIALS AND METHODS

This section describes the metrics, datasets, and preprocessing filters used in this work, which focuses on fundus and mammography. Different algorithms are used to preprocess these types of images to highlight regions of interest. Preprocessing filters reported positively in the literature were selected. Those enhanced images are then used in the subsequent phases of artificial vision systems.

Metrics were implemented using Matlab 2020a. Only the entropy metric is available as a function in Matlab, and the rest of the metrics were coded.

A. Metrics

The metrics described below are applied to two-dimensional images represented as matrixes. The measurement for an entire image is based on partial calculations by blocks without overlapping. Therefore, the first step for computing them is to define a block size and divide the image horizontally and vertically into as many blocks as possible (see Fig. 1). Then, a partial measurement is calculated in each block, and the average of these measurements is considered the value of the metric, taking into account the particularities of each metric.

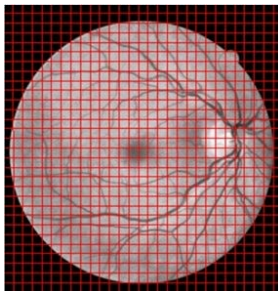


Fig. 1. Example of a fundus medical image divided into square blocks of $L \times L$ size.

1. Enhancement Measure Estimation (EME)

EME is a quantitative measure of contrast enhancement of two-dimensional or grayscale images [21]. The enhancement measure

is a modification of Weber's law and Fechner's law. Weber's law establishes that the visual perception of contrast is independent of luminance and low spatial frequency and determines that the perceived change in intensity is proportional to the initial one. On the other hand, Fechner's law states that perception and stimulation are linked logarithmically (i.e., the visually perceived intensity value is proportional to the logarithm of the actual intensity).

For the calculation of the EME metric, the two-dimensional discrete image of $M \times N$ size is divided into small blocks. M and N are the width and the height of the image in pixel number, respectively; the size of the square block is $L \times L$, for $L = 3, 4, \dots, n$. Then, the minimum and maximum intensity of each block are found, the contribution of each block is calculated as a natural logarithm function. Finally, the EME value for the whole image is equal to the average. EME is mathematically expressed in (1), where $k_1 = M/L$, is the number of horizontal blocks; $k_2 = N/L$, it is the number of vertical blocks; $I_{max}^{l,m}$ y $I_{min}^{l,m}$ are the maximum and minimum pixel intensity values in a block (m, l) , respectively. The size of the block L affects the EME value.

$$EME = \frac{1}{k_1 k_2} \sum_{m=1}^{k_1} \sum_{l=1}^{k_2} 20 \ln \left(\frac{I_{max}^{l,m}}{I_{min}^{l,m}} \right) \quad (1)$$

Note the mathematical equivalence in (2), where $I^{l,m} \in 1, 2, 3, \dots, 256$ corresponds to the intensity values shifted by 1 to avoid the indeterminacy of $\ln(0)$. So that, the computation of the logarithm is possible, $\ln(I^{l,m}) \in \{\ln(1), \ln(2), \dots, \ln(256)\}$. That is, $\ln(I^{l,m}) \in \{0, 0.6931, \dots, 5.5452\}$. Therefore, the difference between the maximum value and the minimum value remains controlled. Additionally, the expression uses the constant 20 to amplify the difference and provide a higher significance value for EME.

$$\ln \left(\frac{I_{max}^{l,m}}{I_{min}^{l,m}} \right) = \ln(I_{max}^{l,m}) - \ln(I_{min}^{l,m}) \quad (2)$$

2. Enhancement Measure Estimation by Entropy (EMEE)

This metric is the entropy measure that relates the contrast for each block, scaled by a parameter (α), and averaged over the entire image [21]. The value of α , for $0 < \alpha < 1$, is proportional to the emphasis of the entropy. The variable α helps to manage more randomness. The calculation formula is shown in (3). The impact of the block size (L) and the entropy emphasis (α) over EMEE calculation will be discussed in the Results section.

$$EMEE = \frac{1}{k_1 k_2} \sum_{m=1}^{k_1} \sum_{l=1}^{k_2} \alpha \left(\frac{I_{max}^{l,m}}{I_{min}^{l,m}} \right)^\alpha \ln \left(\frac{I_{max}^{l,m}}{I_{min}^{l,m}} \right) \quad (3)$$

The factor $\alpha \left(\frac{I_{max}^{l,m}}{I_{min}^{l,m}} \right)^\alpha$ highlights the logarithmic amplitude of $I^{l,m}$, increasing the large and compressing minor ones (See Fig. 2). This measure could be more suitable for images that have greater visible variability.

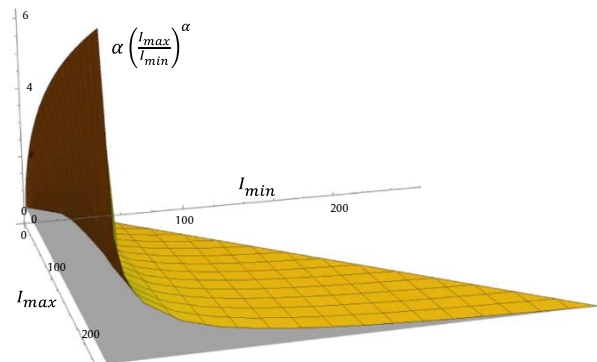


Fig. 2. Weighting factor $\alpha \left(\frac{I_{max}^{l,m}}{I_{min}^{l,m}} \right)^\alpha$, $\alpha = 1/2$, $I_{min} \leq I_{max}$.

3. Entropy

Entropy is a statistical measure of randomness. It could characterize textures from an image or photograph [15], [22]. The texture within an image is an element that holds qualities perceived through sight and touch. Texture depicts aspects of the surfaces of photographed objects or subjects. The entropy is calculated on a grayscale image from its histogram. The entropy formula is defined in (4) and (5).

$$\text{Entropy} = - \sum_{j=0}^{255} E(I_j) \quad (4)$$

$$E(I_j) = \begin{cases} 0 & \text{if } P(I_j) = 0 \\ P(I_j) \log_2 P(I_j) & \text{if } P(I_j) > 0 \end{cases} \quad (5)$$

In the equation (5), $P(I_j) \in [0,1]$ and $P(I_j)$ is the probability of occurrence of the j th intensity, according to the image histogram.

4. Contrast Index CI and Contrast Improvement Index (CII)

The Contrast Index (CI) is a measure that establishes the relationship between the background and the foreground of an image. This ratio is calculated by blocks, and the average among these blocks will be the CI measure of the entire image [4]. A high CI value means a more significant difference between background and foreground, and therefore the image presents a good CI. The contrast measure of an image is expressed in (6).

$$CI = \frac{1}{k_1 k_2} \sum_{m=1}^{k_1} \sum_{l=1}^{k_2} \left(\frac{l_{max}^{lm} - l_{min}^{lm}}{l_{max}^{lm} + l_{min}^{lm}} \right) \quad (6)$$

To establish the improvement of an image, CII is defined in the equation (7). CII is the ratio between the contrast index of the resulting image and the contrast index of the original image. If this ratio is greater than one, the contrast is considered to have been improved.

$$CII = \frac{CI_{processed}}{CI_{original}} \quad (7)$$

B. Fundus Datasets

Fundus images are obtained by examining the retina during pathology inspections. The retinal datasets used in this work contain images of adults and children. Adult datasets are predominant. Although there is a very active scientific community applying machine vision to diagnose child retinal pathologies, there is a lack of public data. Both adults and children may have eye diseases. Some retina diseases are Retinopathy of Prematurity (ROP), Retinopathy of Hypertension, Diabetic Retinopathy, Glaucoma, etc.

1. ROPFI Dataset

Retinopathy of Prematurity Fundus Images (ROPFI) is a set of children's images with Retinopathy of Prematurity [20]. This set entails 64 images captured with a RetCam Shuttle camera (Clarity Medical Systems, Inc.). The average size of an image is 640 x 480 pixels.

2. DRIVE Dataset

Digital Retinal Images for Vessel Extraction (DRIVE) is a public database of adult fundus images [23]. It comprises 40 images whose dimension is 565 x 584 pixels on average. DRIVE is widely used to validate vascular network segmentation algorithms.

3. HRF POOR-QUALITY Dataset

This database contains images of adult patients, and it belongs to a collaborative research group to support comparative studies about automatic segmentation algorithms [24]. It is a set of eighteen images that are of lower quality than others of the same project. The average dimension is 640 x 460 pixels.

4. Masks Acquisition

ROPFI dataset provides the masks that were obtained automatically [20]. Additionally, DRIVE dataset also contains its corresponding

masks [23]. While the HRF POOR-QUALITY dataset does not include masks, they have been computed performing the automatic procedure in [20]. First, an Otsu binarization was performed on the green channel. Then a convolution was applied to find borders, and finally the maximum contour was selected to set the binary mask.

C. Mammogram Datasets

1. MIAS Dataset

The Mammographic Image Analysis Society database (MIAS) is a set of 100 labeled and annotated images [25]. The database is suitable for performing and understanding mammograms' technical and visual analysis for research purposes, such as anomalies detection algorithms and other technological derivations that allow computerized assistance to medical specialists. This mammography database is in PGM format, and it contains one hundred images. The average size of the images is 475 x 933 pixels.

2. DDSM Dataset

The Digital Database for Screening Mammography (DDSM) has been published by the University of Florida [26]. DDSM dataset contains 31 mammograms. This dataset is a resource for the mammographic image analysis research community. The main objective of the database is to facilitate research for the development of computer algorithms to assist in the detection of mammography anomalies. It includes the diagnosis and development of assistance aids through software for medical specialists.

3. Automatic Mask Creation

MIAS and DDSM mammogram datasets do not provide masks. Hence, the automatic procedure based on [20] was applied.

D. Filters for Processing Fundus Images

In both infant and adult fundus images, it may not be possible to distinguish attributes of the retina, such as the presence of vessels. It could cause a misdiagnosis by a medical specialist. In [20], this problem is indicated: childhood retinal images present difficulty recognizing retinal elements because they are low in contrast and brightness, have small, curved lines, and present noise. Thus, the authors propose to apply some filters in sequence: contrast, brightness, gamma correction, and CLAHE.

A general image processing operator is a transformation that takes an input I image and computes an output image G . If the image is color, for example, in the red, green, and blue (RGB) color space, the image is usually split into its channels. After applying the filter to each channel, the image can be reconstructed into a color image from its modified channels [27].

The brightness and contrast transformation of one channel at a time can be expressed as in (8):

$$G(x) = c \times I(x)[h] + b \quad (8)$$

The c and b parameters represent the contrast and brightness values that will modify the image, respectively; $I(x)$ describes the intensity image in a pixel x ; and h denotes the particular red, green, or blue channel.

Then, a reverse gamma correction filter [27] is applied. This filter removes the non-linear schema of the input image that has been acquired with a digital camera and provides a brighter image using a correction value greater than one. The reverse gamma correction filter is mathematically expressed in (9), where γ is the gamma correction parameter. The gamma transformation is applied to each color image channel. The modified channels are combined, and an image in the same color space as the original image $I(x)$ is obtained.

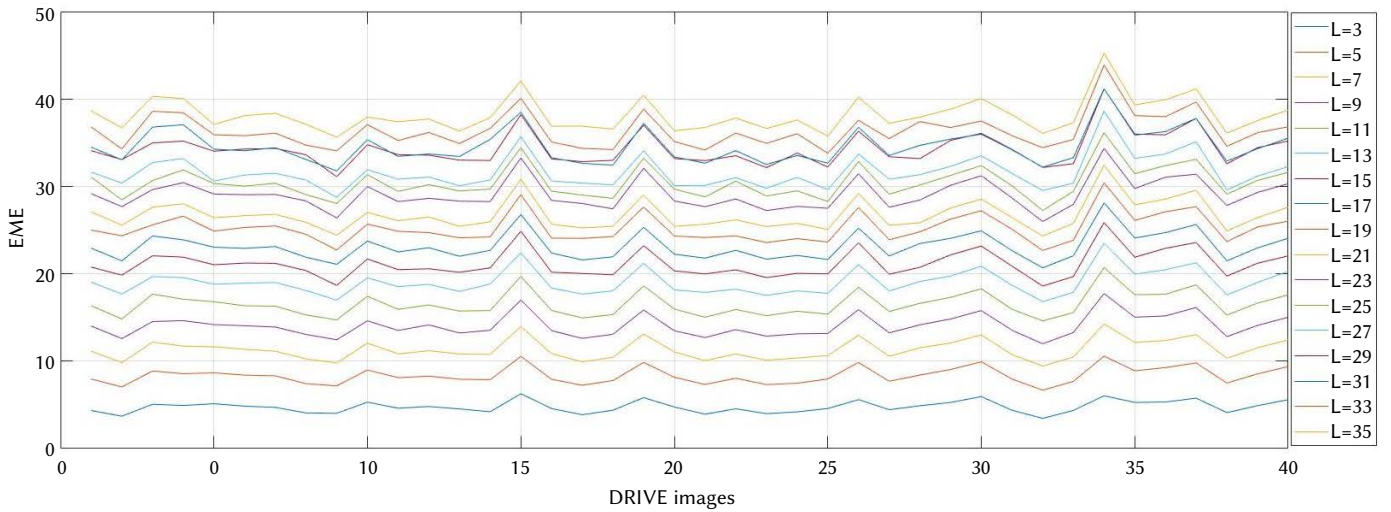


Fig. 3. Computation of EME metric using DRIVE dataset, and varying the block size L . The line color represents a different L value, $L = \{3, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35\}$ from bottom to top.

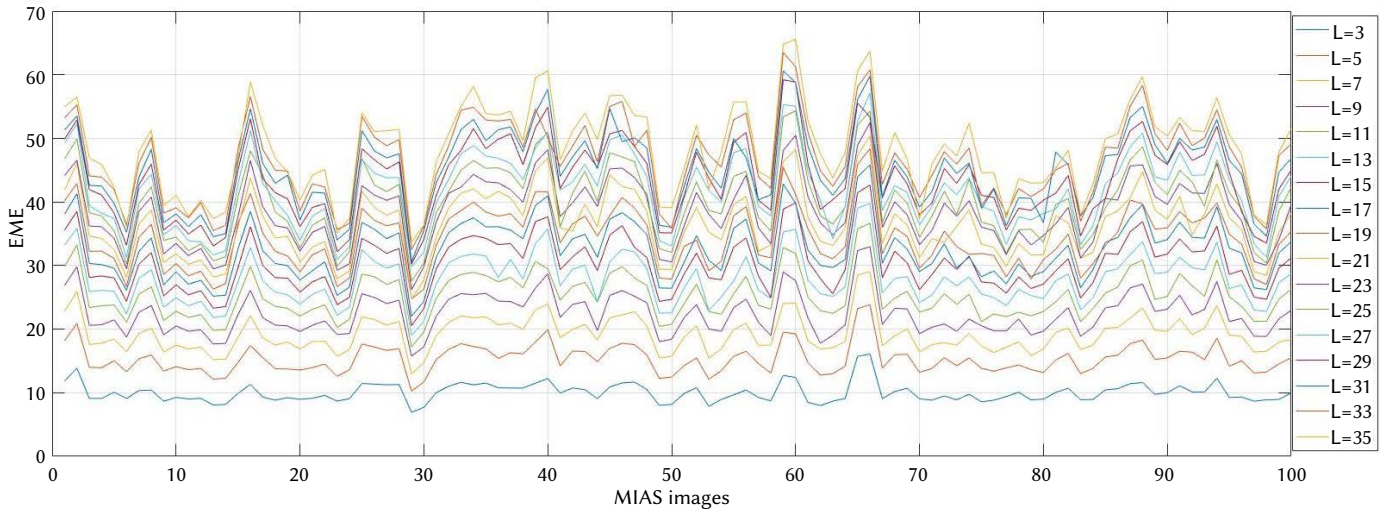


Fig. 4. Computation of EME metric using MIAS dataset, and varying the block size L . The line color represents a different L value, $L = \{3, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35\}$ from bottom to top.

$$G(x) = I(x)[h]^{\frac{1}{r}} \quad (9)$$

The last filter of the sequence proposed by the authors is the CLAHE [28]. This filter is applied to the L channel of the CIE Lab color space. The implemented function to process this filter is shown in expression (10), where k and cl represent the variables kernel size and transformation limit, respectively, and L denotes the luminance channel in the corresponding color space.

$$G(x) = \text{CLAHE}(I(x)[L], k, cl) \quad (10)$$

E. Filters for Processing Mammogram Images

Mammography images are pre-processed with filters CLAHE y Fast Local Laplacian (FLL) in [29]. The CLAHE filter is used to improve mammography contrast, and FLL is employed to reduce noise and smooth the image. The mathematical representation of CLAHE was given in expression (10). The FLL filter is represented mathematically in (11), where u and v are the coordinates over the x and y axis, respectively, and σ is the standard deviation.

$$G_0 = G_\sigma(u, v) \cdot e^{-\frac{u^2 + v^2}{2\sigma^2}} \quad (11)$$

IV. RESULTS

This section begins by analyzing the parameters of each metric and observing how they affect the calculation of the image improvement. Then, the quantitative enhancement of every dataset through each metric is assessed.

A. Analysis of EME Variables

The calculation of EME is based on the size of the block (L). It determines the partial area used to calculate each sum given in (1). EME has been computed for the five datasets shown in the previous section, setting L values from 5 to 19. The analysis of L values is supported by analyzing the plot of the five datasets. As illustrative examples, Fig. 3 and Fig. 4 display the behavior of L in the datasets DRIVE and MIAS, respectively. According to the behavioral analysis of the metric, it has been observed that, with low values of L , the metric reports a low weight as well, which indicates a minor improvement. In contrast, high L values increase the value of the metric. Based on that, and to obtain significant values showing an improvement in the image, a value of L equal to 19 was selected.

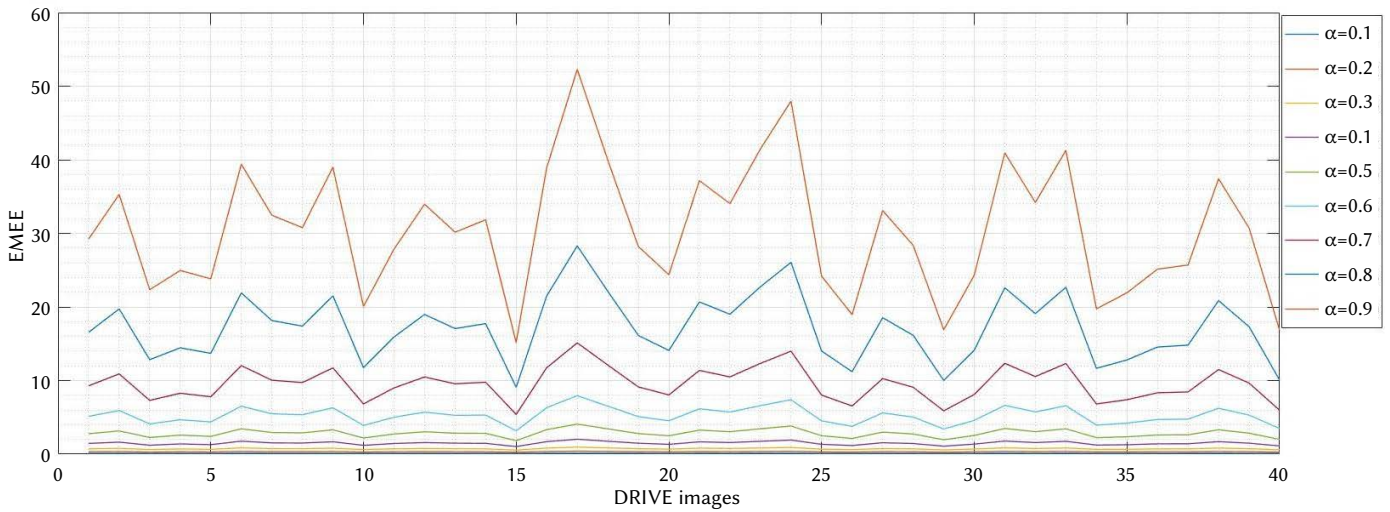


Fig. 5. Computation of EMEE using DRIVE dataset, varying α from 0.1 to 0.9 and keeping block size $L=19$. The line color represents a different α value, $\alpha = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.9\}$ from bottom to top.

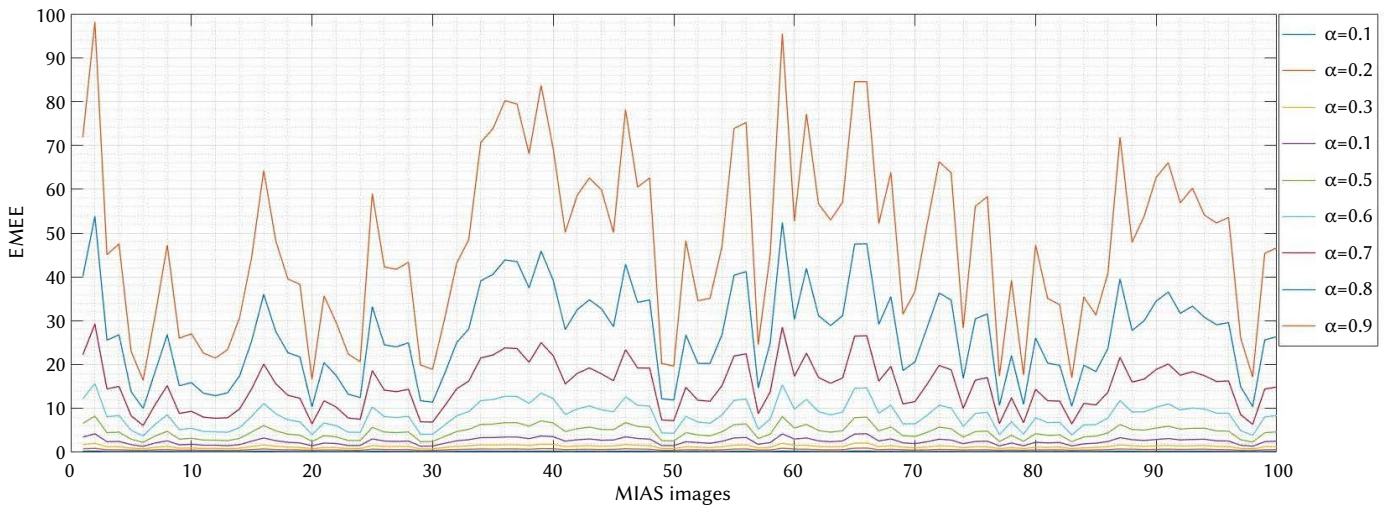


Fig. 6. Computation of EMEE using MIAS dataset, varying α from 0.1 to 0.9 and keeping block size $L=19$. The line color represents a different α value, $\alpha = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.9\}$ from bottom to top.

B. Analysis of EMEE Variables

As mentioned above, the EMEE metric has two variables that will influence the calculation: the entropy effect (α) and the block size (L). The behavior of α between 0.1 and 0.9 is analyzed. L is set to the value of 19, as commented in the previous section.

The influence of the variable α for each dataset was analyzed by calculating and plotting EMEE for each dataset. As illustrative examples, Fig. 5 and Fig. 6 show the EMEE metric using both DRIVE of fundus and MIAS of mammograms datasets, respectively. These plots show EMEE as function of α . The EMEE calculation obtained significant values when alpha was set equal to 0.7, 0.8, and 0.9. Taking a middle point alpha has been chosen equal to 0.8. This value will be used for the calculation of the EMME metric.

C. Evaluation of the Improvement on the Fundus Datasets

Images from DRIVE, ROPFI, and HRF POOR-QUALITY datasets were improved using the enhancement method proposed in [20], and detailed in Subsection III.D. An important point is to use a mask for delimiting the region of interest (ROI). The mask is a binary image with values 0 and 255, where the pixels with an intensity value equal to 255 constitute the ROI. The discarded part of the image is in black.

Fig. 7, Fig. 8, and Fig. 9 show an image example of DRIVE, ROPFI, and HRF POOR-QUALITY, respectively, that have been preprocessed. Original, original in grayscale, mask, and enhanced images are included. The complete datasets were enhanced with the corresponding algorithm, and subsequently images were evaluated with each metric.

Considering that (6) represents CII as the ratio between the improved image and the original image. Similarly, the ratio of the rest of the measures has been computed. These improvement rates on average for the DRIVE, ROPFI, and HRF POOR-QUALITY datasets are shown in Table I.

TABLE I. PERFORMANCE OF THE FUNDUS IMAGES ENHANCEMENT. THE PARAMETERS USED WERE $L = 19$, AND $\alpha = 0.8$. AVERAGE OF ENHANCEMENT (AE) RATE. AVERAGE OF ENHANCEMENT PERCENTAGE (AEP)

Dataset	Performance	CII	EME	EMEE	Entropy
DRIVE	AE Rate	1.18	1.55	2.37	1.02
	AEP (%)	18.11	55.42	137.47	1.95
ROPFI	AE Rate	1.03	1.75	1.34	1.09
	AEP (%)	3.59	75.32	34.40	8.98
HRF POOR-Q	AE Rate	1.4195	1.66	2.86	1.02
	AEP (%)	41.95	66.45	186.40	1.95

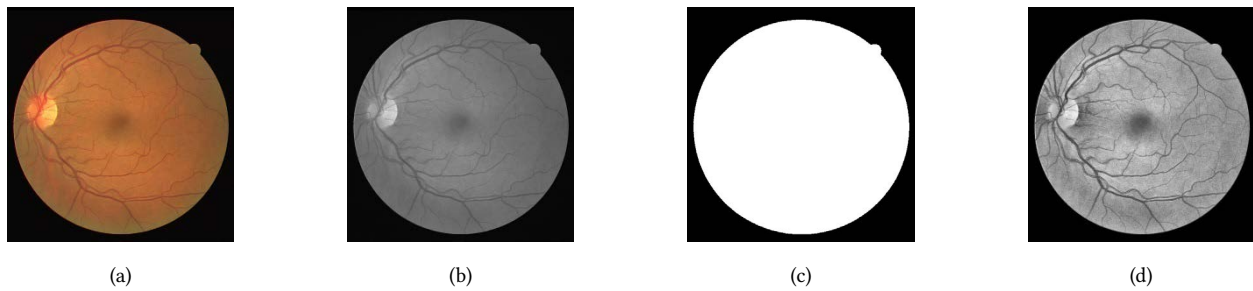


Fig. 7. First image of the DRIVE dataset. (a) original image, (b) original in grayscale, (c) mask, (d) enhanced image.

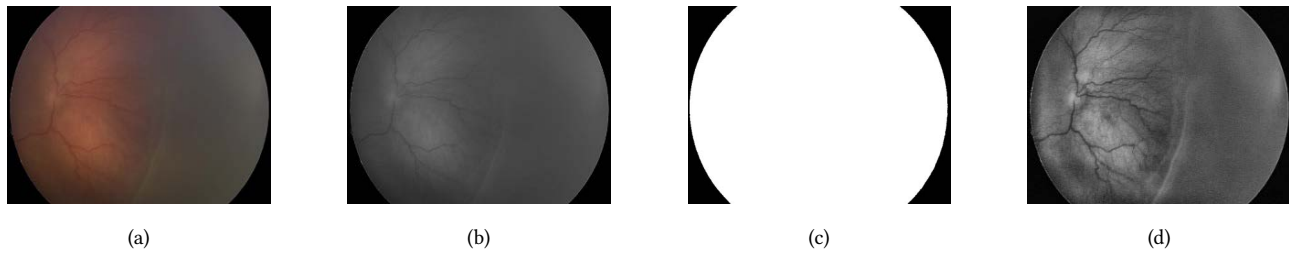


Fig. 8. First image of the ROPFI dataset. (a) original image, (b) original in grayscale, (c) mask, (d) enhanced image.

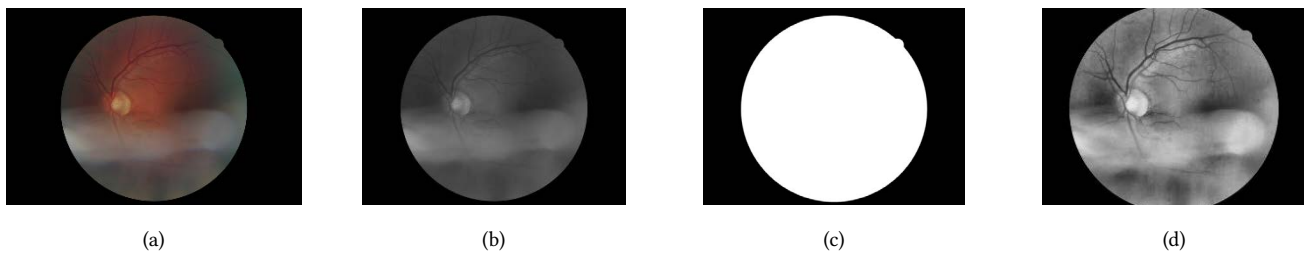


Fig. 9. First image of the HRF POOR-QUALITY dataset. (a) original image, (b) original in grayscale, (c) mask, (d) enhanced image.

The average improvement in the DRIVE dataset has been 137%, 55%, 18%, and 2%, applying EMEE, EME, CII, and entropy, respectively. Looking at the report of each metric and considering the averages in Table I, it is possible to conclude that the metrics that achieve the best quantitative distinction are EMEE and EME.

The ROPFI dataset's average improvement percentages reported by EME, EMEE, entropy, and CII have achieved 75%, 64%, 9%, and 4%, respectively. Looking at the report of each metric in Table 1, the metrics that achieve the best quantitative distinction are EME and EMEE.

Concerning the HRF POOR-QUALITY dataset, the improvement percentages on average of 186% and 66% of EME, EMEE, respectively, have achieved the best quantitative distinction.

D. Evaluation of the Improvement on the Mammogram Datasets

As previously mentioned, the MIAS and the DDSM datasets are mammography images. Using the preprocessing method proposed in [27], these sets were improved and commented in Subsection III.E.

Analogously to the evaluation of the fundus images, original and mask images are employed. Each mammography has been pre-processed, and the contrast measurement has been calculated in both the original and preprocessed images. The contrast measurement is only computed for the region of interest using the mask image. The contrast measurement on average for the original and improved images has been calculated, and, finally, the improvement ratio has been obtained. The rate and percentage enhancement on average are stated in Table II.

TABLE II. ENHANCEMENT PERFORMANCE OF MAMMOGRAPHY DATASETS. THE PARAMETERS USED WERE $L = 19$, AND $\alpha = 0.8$. AVERAGE OF ENHANCEMENT (AE) RATE. AVERAGE OF ENHANCEMENT PERCENTAGE (AEP)

Dataset	Performance	CII	EME	EMEE	Entropy
MIAS	AE Rate	4.96	2.70	4.53	1.06
	AEP (%)	396.55	170.11	353.39	6.05
DDSM	AE Rate	2.30	1.57	1.82	1.04
	AEP (%)	129.92	56.97	82.04	4.16

A pair of improved images of the MIAS and DDSM datasets are presented in Fig. 10 and 11, respectively. These figures incorporate original in color, original in grayscale, mask, and preprocessed, respectively.

In the case of the MIAS dataset, CII and EMEE indicated the highest enhancement values, 396% and 353% on average, respectively. EME also presents a high average value with respect to the Entropy average, 160%, and 2%, respectively.

With respect to the DDSM dataset, CII is the most significant value, which reported 129% on enhancement average. EMEE, EME, and entropy registered 66%, 57%, and 4% on enhancement average, respectively. Thus, EMEE is close to the EME value, and both are quite large with respect to the Entropy.

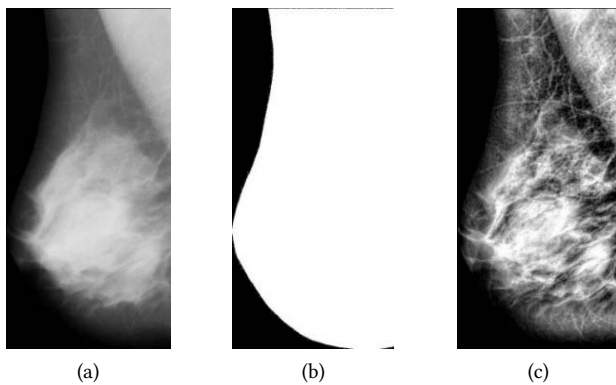


Fig. 10. First image of the MIAS dataset. (a) original image, (b) mask, (c) enhanced image.

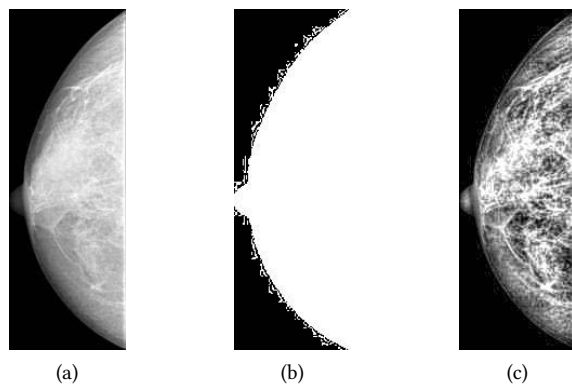


Fig. 11. First image of the DDSM dataset. (a) original image, (b) mask, (c) enhanced image.

V. DISCUSSION

In a computer vision system, contrast and brightness enhancement of images is part of the first phase named preprocessing, and it is essential for the following phases. This research deals with fundus images of the retina and mammograms, including healthy and pathological cases. The improvement algorithms could be assessed through experts' criteria and quantitative validation. Expert criteria agreement may require the analysis of several experts and could be subjective.

In the literature review, it has been commented that it is not common in artificial vision systems to report the image improvement quality using quantitative metrics. It seems that scientists rely on visual perception or on trial and error. On the other hand, there have been a few works that have been considered quantitative aspects, as discussed in previous sections.

Subsequent phases of the computer vision system may receive an image with a non-satisfactory enhancement, and brightness and contrast alteration may need to be performed again. Consequently, appropriate quantitative measures are quite valuable.

In this research, medical images of retina fundus and mammograms have been selected to study and quantify contrast measurements. These two types of pathological images are of broad interest to physicians and informatics specialists. Because fundus images usually are colored, and mammograms are grayscale, these datasets permit to evaluate the proposed measures in both colored and grayscale images. Regarding the similarities, it can be observed that both types examine anatomical parts of the human body, and that the background is a large portion of the image (so that binary masks are used to delimit the area of interest). Because of these common characteristics, it was possible to examine both datasets using similar scripts.

Previous works agree that a high metric value represents a better distinction of the parts of interest versus the background [2], [6], [12]. However, the parameter's values were not reported. Due to this, the first effort of this work has been to establish the most suitable parameters of each metric through mathematical analysis and experimentation. Those most suitable parameters have been identified as the block size ($L = 19$), and the entropy emphasis, ($\alpha = 0.8$), as treated in the Results Section.

In our case studies, in mammograms, CII and EMEE reported the highest contrast enhancement rates of up to 396% and 353%, respectively, and EME of up to 170%; and, regarding retinal datasets, EMEE, CII, and EME metrics reported enhancements of up to 186%, 75%, and 41%, respectively. Entropy is the measure with the smallest margin of distinction in both fundus and mammography images. However, there is a high improvement ratio in the case of mammograms compared to fundus images. The improvement percentage of the entropy metric ranges from 2% to 8%. Accordingly, it can be recommended using the EMEE, EME and CII metrics to quantitatively validate the contrast and brightness improvement of medical images.

An analysis of the behavior of the measurement as a function of parameters L and α was carried out. This being so, the chosen parameters allow differentiating better the image improvement. The values of each parameter have been studied and reported precisely, with the intention that researchers who need to use those metrics know the most convenient parameters.

In Section I, it was commented that the image enhancement algorithm influences and improves, in most cases, the performance of the artificial vision system. We have presented how the metrics can achieve more significant improvements in certain types of images. Therefore, for images similar to the cases studied in the paper, the preprocessing and metric values presented could be applied. For other, quite different images, the analysis and guidelines presented in the paper can be adapted to perform the analysis and parameter selection.

The scope of this work has been to study evaluation metrics of image enhancement algorithms. Also, this work has been considered mammography and fundus images. In order to include various types of healthy and pathological images, three different fundus datasets and two different mammography datasets were included.

This research work could be most valuable for researchers that develop computer vision applications, in order to evaluate the quality of their preprocessed images and improve the applicability of their techniques.

Since the amount and variety of datasets have not been extensive, the main future works are to extend this research by evaluating other sets of related medical images, reproduce a complete computer vision method, and report the relation between quantitative enhancement and the computer vision system performance.

VI. CONCLUSION

The review of related works indicated that the image preprocessing phase affects the results achieved by subsequent steps of an artificial vision system. As reported, the correct preprocessing of the input images accomplished that deep neural network techniques could improve up to 4% their accuracy. Thus, the consideration of this early quantitative assessment of image quality could be incorporated into the design of machine vision systems in the medical imaging field.

The need to quantitatively validate the enhancement of medical images in the first phase (or preprocessing) of a computer vision system was a main motivation of this research work. And, as discussed in the paper, metrics EMEE, EME and CII are valuable for measuring the enhancement of the studied medical images.

To apply these metrics in new datasets, an analysis of the metrics parameters following the approach of this paper is recommended. An important consideration is that the region of interest of images should be satisfactorily delimited.

In future work, it is planned to initiate a collaboration with additional clinical specialists to gather their opinions and suggestions about the preprocessing phase, so that they could be taken into account in future developments.

ACKNOWLEDGMENT

We thank Metrofraternidad Foundation, Quito, Ecuador, for providing the ROPFI dataset.

REFERENCES

- [1] Y. Boutiche, "Fast level set algorithm for extraction and evaluation of weld defects in radiographic images," *Studies in Computational Intelligence*, vol. 672, pp. 51–68, 2017, doi: 10.1007/978-3-319-46245-5_4.
- [2] X. Chen, "Image enhancement effect on the performance of convolutional neural networks", M.S. thesis, Faculty of Computing, Blekinge Institute of Technology, Karlskrona, Sweden, 2019, doi:10.1016/j.compbimed.2021.104319.
- [3] T. Rahman, et. al, "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images", *Computers in Biology and Medicine*, vol. 132, 2021.
- [4] H. Qu, T. Yuan, Z. Sheng, and Y. Zhang, "A Pedestrian Detection Method Based on YOLOv3 Model and Image Enhanced by Retinex," *Proceedings - 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI 2018*, Feb. 2019, doi: 10.1109/CISP-BMEI.2018.8633119.
- [5] D. A. Pitaloka, A. Wulandari, T. Basaruddin, and D. Y. Liliana, "Enhancing CNN with Preprocessing Stage in Automatic Emotion Recognition," *Procedia Computer Science*, vol. 116, pp. 523–529, Jan. 2017, doi: 10.1016/J.PROCS.2017.10.038.
- [6] E. Vocaturo, E. Zumpano and P. Veltri, "Image pre-processing in computer vision systems for melanoma detection," 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018, pp. 2117–2124, doi: 10.1109/BIBM.2018.8621507.
- [7] J. P. Gu, L. Hua, X. Wu, H. Yang, and Z. T. Zhou, "Color medical image enhancement based on adaptive equalization of intensity numbers matrix histogram," *International Journal of Automation and Computing*, vol. 12, no. 5, pp. 551–558, 2015, doi: 10.1007/s11633-014-0871-9.
- [8] M. M. G. Ribeiro, and A. J. P. Gomes, "RGBeat: A Recoloring Algorithm for Deutan and Protan Dichromats," *International Journal of Interactive Multimedia and Artificial Intelligence*, In Press, pp. 1–13, 2022, doi: 10.9781/ijimai.2022.01.003.
- [9] D. R. Flatla, K. Reinecke, C. Gutwin, and K. Z. Gajos, "SPRWeb: preserving subjective responses to website colour schemes through automatic recolouring," in *Proc. Conf. Human Factors in Computing Systems (SIGCHI'13)*. ACM, 2013, pp. 2069–2078, doi: 10.1145/2470654.2481283.
- [10] X. Xu, Y. Wang, J. Tang, X. Zhang, and X. Liu, "Robust automatic focus algorithm for low contrast images using a new contrast measure," *Sensors*, vol. 11, no. 9, pp. 8281–8294, 2011, doi: 10.3390/s110908281.
- [11] S. Chen, C. Wang, I. Tai, K. W. Y. Chen, and K. Hsieh, "Modified YOLOv4-DenseNet Algorithm for Detection of Ventricular Septal Defects in Ultrasound Images," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 7, pp. 101–108, 2021, doi: 10.9781/ijimai.2021.06.001.
- [12] S. Wu, Q. Zhu, Y. Yang, and Y. Xie, "Feature and contrast enhancement of mammographic image based on multiscale analysis and morphology," *2013 IEEE International Conference on Information and Automation, ICA 2013*, vol. 2013, pp. 521–526, 2013, doi: 10.1109/ICInfA.2013.6720354.
- [13] A. Pandey and S. Singh, "New performance metric for quantitative evaluation of enhancement in mammograms," *Proceedings of the 2013 2nd International Conference on Information Management in the Knowledge Economy, IMKE 2013*, pp. 51–56, 2014.
- [14] G. Du et al., "A new method for detecting architectural distortion in mammograms by nonsubsampling contourlet transform and improved PCNN," *Applied Sciences (Switzerland)*, vol. 9, no. 22, 2019, doi: 10.3390/app9224916.
- [15] S. Gupta and R. Porwal, "Appropriate Contrast Enhancement Measures for Brain and Breast Cancer Images," *International Journal of Biomedical Imaging*, vol. 2016, no. 1, 2016, doi: 10.1155/2016/4710842.
- [16] A. Albiol, A. Corbi, and F. Albiol, "Automatic intensity windowing of mammographic images based on a perceptual metric," *Medical Physics*, vol. 44, no. 4, pp. 1369–1378, Apr. 2017, doi: 10.1002/MP.12144.
- [17] A. Albiol, A. Corbi, and F. Albiol, "Measuring X-ray image quality using a perceptual metric," *2016 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges, GMEPE/PAHCE 2016*, Jul. 2016, doi: 10.1109/GMEPE-PAHCE.2016.7504639.
- [18] K. Aurangzeb, S. Aslam, M. Alhussain, R. A. Naqvi, M. Arsalan, and S. I. Haider, "Contrast Enhancement of Fundus Images by Employing Modified PSO for Improving the Performance of Deep Learning Models," *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3068477.
- [19] K. L. Nisha, S. G., P. S. Sathidevi, P. Mohanachandran, and A. Vinekar, "A computer-aided diagnosis system for plus disease in retinopathy of prematurity with structure adaptive segmentation and vessel based features," *Computerized Medical Imaging and Graphics*, vol. 74, pp. 72–94, 2019, doi: 10.1016/j.compmedimag.2019.04.003.
- [20] M. Intriago-Pazmino, J. Ibarra-Fiallo, J. Crespo, and R. Alonso-Calvo, "Enhancing vessel visibility in fundus images to aid the diagnosis of retinopathy of prematurity," *Health Informatics Journal*, pp. 1–15, 2020, doi: 10.1177/1460458220935369.
- [21] S. S. Agaian, K. P. Lentz, A. M. Grigoryan, S. S. Agaian, K. P. Lentz, and A. M. Grigoryan, "A New Measure of Image Enhancement," *IATED International Conference on Signal Processing & Communication*, no. January 2000, pp. 19–22, 2000.
- [22] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. New York: Pearson, 2018.
- [23] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-Based Vessel Segmentation in Color Images of the Retina," *IEEE Transactions on Medical Imaging*, vol. 23, no. 4, pp. 501–509, Apr. 2004, doi: 10.1109/TMI.2004.825627.
- [24] A. Budai, R. Bock, A. Maier, J. Hornegger, and G. Michelson, "Robust vessel segmentation in fundus images," *International Journal of Biomedical Imaging*, vol. 2013, 2013, doi: 10.1155/2013/154860.
- [25] J. Suckling, J. Parker, J. Dance, D. Astley, S. Hutt, I. Boggis, C. Ricketts, I. Stamatakis, E. Cerneaz, N. Kok, S. Taylor, P. Betal, D. Savage, "The Mammographic Image Analysis Society Digital Mammogram Database," *Kluwer Academic Publishers*, vol. 13, pp. 375–378, 1994.
- [26] K. Bowyer, K., Kopans, D., Kegelmeyer, W. P., Moore, R., Sallam, M., Chang, K., Woods, "HeathEtAl2001_DDSMdataset," in *In Third international workshop on digital mammography*, 1996, p. 27.
- [27] R. Szeliski, *Image formation*, vol. 73. 2011. doi: 10.1007/978-3-642-2014
- [28] K. Zuiderveld, *Contrast Limited Adaptive Histogram Equalization*. Academic Press, Inc., 1994. doi: 10.1016/b978-0-12-336156-1.50061-6.
- [29] N. Arora and G. Aggarwal, "Mammogram Classification Using Genetic Algorithm-based Feature Selection," *International Journal of Research in Electronics and Computer Engineering a Unit of I2OR*, vol. 6, no. 4, pp. 378–379, 2018.



Monserrate Intriago-Pazmiño

She received the B.S. degree in Computer Science Engineering from National Polytechnic School, Quito, Ecuador, in 2007, and the M.S degree in Computer Science from the Technical University of Madrid, Madrid, Spain, in 2012. She is currently an Associate Professor at the Department of Informatics and Computer Science, at the National Polytechnic School. She is a Ph.D. candidate in Computer Science at Technical University of Madrid. Her research interests include software quality and machine learning.



Julio Ibarra-Fiallo

He received the degree of Mathematician from the National Polytechnic School, Ecuador, and the master's degree in Applied Mathematics, from the San Francisco University of Quito, Ecuador. He is currently an Associate Research Professor at the College of Sciences and Engineering at the San Francisco University of Quito. He has participated and directed national research projects. He researches in areas of artificial intelligence, pattern recognition, computer vision, typical testors, combinatorial logic models of variable selection, neural networks, among others.



Adán Guzmán-Castillo

He received the Engineering degree in Computer and Information Systems from the Escuela Politécnica Nacional (EPN), Ecuador, in 2020. He is currently pursuing the M.Sc. degree at the Escuela Politécnica Nacional (EPN). He is also a Research Technician with EPN. His current research interests include studies of medical image segmentation algorithms and studies of interoperability solutions in IoT environments involving multiple devices working with different technologies.



Raúl Alonso-Calvo

He is a PhD in Computer Science from the Universidad Politécnica de Madrid (UPM). He has been a visitor researcher at Universidade de Aveiro DETI-IEETA (Portugal) (2010), and Oxford University (UK) (2014). Since 2010 he is at the Departamento de Lenguajes Sistemas Informáticos e Ingeniería de Software, ETSI Informáticos at UPM, where he is currently an Associate Professor. He has been a member of the Biomedical Informatics Group at UPM from 2001. His research interests are mainly focused on clinical research informatics, biomedical interoperability standards, database integration and preprocessing, biomedical image processing and information retrieval in biomedicine. He has been author and co-author of research papers in several journals. He has participated in EU research projects since 2001 and in recent years he was involved in INTEGRATE: Driving Excellence in Integrative Cancer and EURECA: Enabling information re-use by linking clinical Research and Care.



José Crespo

He received the degrees of Master of Science (School of Electrical and Computer Engineering), Master of Science (School of Management), and the Ph.D. degree (School of Electrical and Computer Engineering) from the Georgia Institute of Technology (Atlanta, USA). He received the degree of "Ingeniero de Telecomunicaciones" from "Universidad Politécnica de Madrid" (Spain). He is a full Professor at "ETS Ingenieros Informáticos", "Universidad Politécnica de Madrid" (Spain), where he belongs to the Biomedical Informatics Group. He has participated in national and international research projects. He has been Head of Department.

S-Divergence-Based Internal Clustering Validation Index

Krishna Kumar Sharma¹, Ayan Seal^{2,6*}, Anis Yazidi^{3,4,5}, Ondrej Krejcar^{6,7}

¹ Department of Computer Science and Informatics, University of Kota, Kota, Rajasthan-324005 (India)

² Department of Computer Science and Engineering, PDPM Indian Institute of Information Technology Design & Manufacturing Jabalpur, Jabalpur, Madhya Pradesh-482005 (India)

³ Department of Computer Science, OsloMet–Oslo Metropolitan University, Oslo, 460167, (Norway)

⁴ Department of Computer Science, Norwegian University of Science and Technology, Trondheim, 460167, (Norway)

⁵ Department of Plastic and Reconstructive Surgery, Oslo University Hospital, Oslo, 460167, (Norway)

⁶ Center for Basic and Applied Science, Faculty of informatics and management, University of Hradec Kralove, Rokitanskeho 62, 50003 Hradec Kralove, (Czech Republic)

⁷ Malaysia-Japan International Institute of Technology (MJIT), Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, (Malaysia)

Received 22 July 2022 | Accepted 14 January 2023 | Published 24 October 2023



ABSTRACT

A clustering validation index (CVI) is employed to evaluate an algorithm's clustering results. Generally, CVI statistics can be split into three classes, namely internal, external, and relative cluster validations. Most of the existing internal CVIs were designed based on compactness (CM) and separation (SM). The distance between cluster centers is calculated by SM, whereas the CM measures the variance of the cluster. However, the SM between groups is not always captured accurately in highly overlapping classes. In this article, we devise a novel internal CVI that can be regarded as a complementary measure to the landscape of available internal CVIs. Initially, a database's clusters are modeled as a non-parametric density function estimated using kernel density estimation. Then the S-divergence (SD) and S-distance are introduced for measuring the SM and the CM, respectively. The SD is defined based on the concept of Hermitian positive definite matrices applied to density functions. The proposed internal CVI (PM) is the ratio of CM to SM. The PM outperforms the legacy measures presented in the literature on both superficial and realistic databases in various scenarios, according to empirical results from four popular clustering algorithms, including fuzzy k-means, spectral clustering, density peak clustering, and density-based spatial clustering applied to noisy data.

KEYWORDS

Cluster Validity Index, Generalized Mean, K -nearest Neighbors, S-distance, S-divergence, Spectral Clustering, Symmetry Favored.

DOI: 10.9781/ijimai.2023.10.001

I. INTRODUCTION

CLUSTERING is an unsupervised methodology for analyzing a set of data objects by dividing them into subsets such that each group contains similar objects while dissimilar ones end up in different groups [1]–[5]. Thus, the objective of clustering is to mine the data to explore multi-dimensional obscure patterns and hidden structures in the data. Nowadays, clustering has received a great deal of attention among the community of researchers in the area of pattern recognition by the virtue of remarkable academic and commercial applications spanning over a wide range which includes identifying fake news [6], spam filtering [7], market segmentation [8], [9], classifying network traffic [10], detecting fraudulent or criminal activity [11], [12], cybersecurity [13], document analysis [14], drug discovery [15], information retrieval [16], and many more [17]–[22].

A fundamental question in clustering is how to assess the “goodness” of the resulting clusters. The answer to this question is not obvious as it is difficult to devise criteria that determine the optimal partitioning of the data objects into clusters. Obtaining insights about the goodness of clusters using some visualization tools is not a feasible solution when the number of dimensions increases, as human eyes are not accustomed to higher-dimensional spaces. The process of assessing the performance of the clustering algorithm is referred to as cluster validation. According to the clustering validation procedure, the outcome of the clustering phase is validated quantitatively by a Clustering Validation Index (CVI). A CVI can be considered a function that, for a given clustering scheme and database, produces some value that represents the quality of the clustering scheme [23], [24]. In other words, a CVI provides some insight into the quality of grouping. Internal, external, and relative are the three main categories of CVIs. Internal CVIs rely only on the internal information of a given database. Unlike internal CVIs, external CVIs assess the “goodness” of a clustering structure based on provided class labels as external inputs [25]–[28]. On the other hand, relative CVIs evaluate the clustering

* Corresponding author.

E-mail address: ayan@iitdmj.ac.in

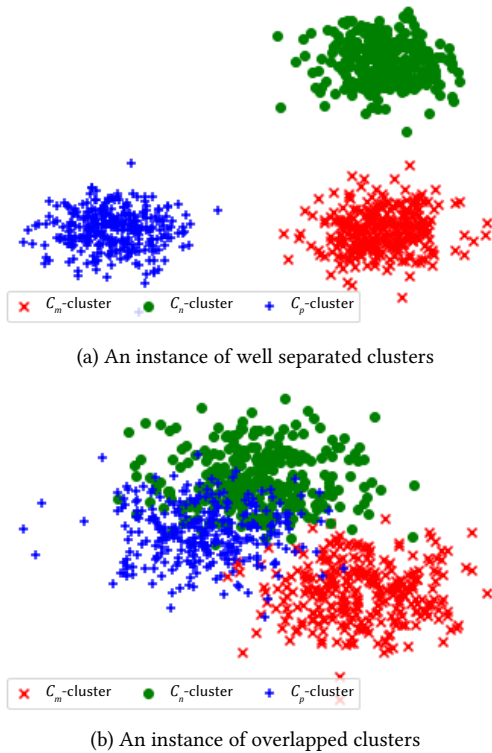


Fig. 1. Distribution of three clusters.

results by changing the number of clusters. We mainly concentrate on internal CVIs in this work. The most intuitive notions for defining “good clustering” are cohesion/compactness (CM) and separation (SM). In simple words, when data objects in a cluster are in the vicinity of each other, the cluster is called a compact cluster. On the other hand, when neighboring clusters are possibly quite far from each other, then these clusters are easily identifiable and well separated. In other words, SM measures the distance between the centers of two clusters, whereas CM measures the variance within a cluster. Generally, geometric distance is used to compute SM. However, geometric distance can not always represent the SM efficiently, especially when two clusters are highly overlapping. Let us consider an example where three clusters, namely C_m , C_n , and C_p , are well separated (see Fig. 1(a)). As clusters are well separated, geometric distance can efficiently capture the dissimilarity between clusters. We may assume another scenario (see Fig. 1(b)), where clusters C_m , C_n , and C_p , are overlapping. In this case, the geometric distance between the centers of C_m and C_n is the same as the geometric distance between the centers of C_m and C_p . Thus, the dissimilarity between clusters can not be captured accurately using geometric distance. In [29], Cui et al. assumed that the data of a cluster were obtained from multivariate Gaussian distributions, and Jeffrey divergence (JD) was considered, as a distance measure for computing SM between clusters. The JD is not a valid distance measure because it does not abide by the metric property of triangle inequality [30]. In addition, the JD is not appropriate, while clusters are almost identical. Alternatively stated, a small change in clusters cannot be captured by JD. It encourages us to delve further in this direction by proposing an internal CVI based on the notion of S-divergence (SD), which can catch tiny variations in clusters since the cone is formed by the Hermitian positive definite matrices (HPDM). In addition, it fulfills all the properties of the distance metric [31]. Four well-known clustering techniques are employed to evaluate the performance of the proposed CVI on ten real-world and artificial databases. However, each cluster is modeled as a random variable using a non-parametric probability density function named kernel density estimation (KDE)

before the use of the proposed internal CVI. Among ten databases, some are well separated, a few are slightly overlapping, and the rest are highly overlapping. Moreover, noise is added in some databases to validate the efficacy of the proposed CVI. A comparative analysis is also performed to show the competitiveness of the proposed internal CVI in comparison with other CVIs.

The remaining article is structured as follows. After the introduction, in Section II, we examine several well-known internal CVIs. The proposed internal CVI is discussed further in Section III. Section IV provides the results of the experiments. At last, Section V concludes the work.

II. RELATED WORKS

A summary of some of the most popular internal CVIs is presented in this section. The CM reflects the average closeness or similarity of data points in all clusters. A value approaching 0 indicates good clustering [32]. The SM portrays the degree of separation between clusters [32]. A higher value of SM signifies better clustering. It is worth mentioning that other indexes, for example, root mean square standard deviation index (RMSSTDI) [33], root squared index (RSI) [33], and modified Hubert validity index (MHI) [34] perform on a different principle. Indeed, the RMSSTDI quantifies the homogeneity of the resultant clusters by calculating the square root of the aggregated variance of all the data objects. RSI determines the magnitude of difference between clusters using the ratio of the addition of the squares **between-clusters** to the total summation of the squares in the database. RMSSTDI, RSI, and MHI evaluate the difference **between-clusters** by calculating the disagreements of groups of data objects in two parts. Furthermore, these indexes do not consider both CM and SM to validate the formed clusters. The Calinski-Harabasz index (CHI) computes the ratio of the sum of the average of **between-clusters** and of **intra-cluster dispersion** for all clusters [35]. A greater value of CHI demonstrates better partitions. CHI is usually fast to compute. Moreover, it is suitable for convex and well-separated clusters. On the other hand, it produces a low value for non-convex clusters. The Dunn validity index (DVI) calculates the SM of clusters over the CM of clusters [36]. Thus, a larger value of DVI suggests well-separated and compact clusters. However, the complexity of the DVI increases with the increase in the number of clusters, k . The Davies-Bouldin index (DBI) computes cluster overlapping using the ratio of the sum of **intra-cluster spread** to **between-cluster distance** [37]. A value adjacent to 0 illustrates better partitions. It computes the inherent attributes and quantities of a database. Moreover, it is limited to Euclidean space. The JD-based validity index (JI) is a ratio of CM to JD-based SM, [38]. JD determines the similarity between two probability distributions and is suitable for slightly-overlapping clusters. Thus, a value close to 0 is a sign of better partitions. However, JD falls short when the clusters are highly overlapping. The silhouette index (SI) measures how alike a data object is to its own cluster/cohesion/CM against other clusters/SM [39]. A value near 1 signifies that the data object is well-suited to its cluster and does not match enough to neighboring clusters. A clustering configuration is appropriate when most data objects have a high value. SI is higher for well-separated and dense clusters. However, it is not suitable for non-convex clusters. Moreover, the computational complexity, $O(n^2d \log(n))$, is high. I validity index (IVI) computes the CM and the SM using the maximum distance among data objects and centers of clusters [40]. Furthermore, the optimal number of clusters is calculated by maximizing the value of IVI. The Xie-Beni index (XBI) is defined using CM as the mean square distance among data objects and their cluster centers and the SM as the minimum square distance between the centers of clusters [41]. Optimal clusters exhibit a minimum value of XBI. The value of XBI reduces monotonically as the value of k increases. Furthermore, Bouguessa et al. [42] and Arbelaitz

TABLE I. A REVIEW OF SOME OF THE POPULAR INTERNAL CVIS

S. No.	Internal CVI	Notation	Expression	Range	Optimal value	Complexity
1	Root mean square standard deviation index	RMSSTDI	$\left\{ \frac{\sum_{i=1}^k \sum_{c_j \in C_i} \ c_j - v_i\ ^2}{d \sum_{i=1}^k (C_i - 1)} \right\}^{1/2}$	[0, +∞]	elbow	$O(nd)$
2	Root squared index	RSI	$\frac{\sum_{c \in DB} \ c - v\ ^2 - \sum_{i=1}^k \sum_{c_j \in C_i} \ c_j - v_i\ ^2}{\sum_{c \in DB} \ c - v\ ^2}$	[0, 1]	elbow	$O(nd)$
3	Modified Hubert validity index	MHI	$\frac{2}{n(n-1)} \sum_{c_j \in C_i \text{ and } v_i \in C_i} \sum_{c_q \in C_r \text{ and } v_r \in C_r} \text{dist}(c_j, c_q) \text{dist}(v_i, v_r)$	[0, +∞]	elbow	$O(n^2 d)$
4	Compactness measure	CM	$\frac{1}{k} \sum_{i=1}^k \frac{1}{ C_i } \sum_{c_j \in C_i} \text{dist}(c_j, v_i)$	[0, +∞]	Min	$O(nd)$
5	Separation measure	SM	$\frac{2}{k^2 - k} \sum_{i=1}^k \sum_{p=i+1}^k \text{dist}(v_i, v_p)$	[0, +∞]	Max	$O(k^2 d)$
6	Calinski-Harabasz index	CHI	$\sum_{i=1}^k \frac{ C_i \times \frac{\text{dist}(v_i, v)}{(k-1)}}{\sum_{c_j \in C_i} \frac{\text{dist}(c_j, v_i)}{(n-k)}}$	[0, +∞]	Max	$O(nd)$
7	Dunn validity index	DVI	$\frac{\min_{1 \leq i \neq j \leq k} \left(\min_{v_{c_{i_0} \in C_i}, v_{c_{j_0} \in C_j}} \{ \text{dist}(c_j, c_q) \} \right)}{\forall c_{c_j} \in C_{i,j}, \forall c_{c_q} \in C_r}$	[0, +∞]	Max	$O(n^2 d \log(n))$
8	Davies-Bouldin index	DBI	$\frac{1}{k} \sum_{i=1}^k \max_{r \neq i} \left(\frac{\frac{1}{ C_i } \sum_{c_j \in C_i} \text{dist}(c_j, v_i) + \frac{1}{ C_r } \sum_{c_j \in C_r} \text{dist}(c_j, v_r)}{\text{dist}(v_i, v_r)} \right)$	[0, +∞]	Min	$O(n^2 d \log(n))$
9	Jeffrey-divergence based validity index	JJ	$\frac{\frac{1}{k} \sum_{i=1}^k \frac{1}{ C_i } \sum_{c_j \in C_i} \text{dist}(c_j, v_i)}{\frac{2}{k^2 - k} \sum_{i=1}^k \sum_{p=i+1}^k JD(v_i, v_p)}$	[0, +∞]	Min	$O(nd)$
10	Silhouette index	SI	$\frac{\frac{1}{n} \sum_{i=1}^k \sum_{c_j \in C_i} \frac{\text{sep}(c_j, c_q) - \text{coh}(c_j, c_i)}{\max\{\text{sep}(c_j, c_q), \text{coh}(c_j, c_i)\}}}{\frac{1}{ C_i } \sum_{c_j \in C_i} \text{dist}(c_j, c_i) \text{ and } \text{sep}(c_j, c_q) = \min_{c_r \neq i \text{ and } 1 \leq r \leq k} \frac{1}{ C_r } \sum_{c_q \in C_r} \text{dist}(c_j, c_q)}$, where $\text{coh}(c_j, c_i) =$	[-1, 1]	Max	$O(n^2 d \log(n))$
11	I validity index	IVI	$\left(\frac{1}{k} \times \frac{\sum_{c \in DB} \text{dist}(c, v)}{\sum_{i=1}^k \sum_{c_j \in C_i} \text{dist}(c_j, v_i)} \times \max_{1 \leq j \neq i \leq n} \{ \text{dist}(c_j, c_q) \} \right)^d$	[0, +∞]	Max	$O(n^2 d \log(n))$
12	Xie-Beni index	XBI	$\sum_{i=1}^k \frac{\sum_{c_j \in C_i} \text{dist}^2(c_j, v_i)}{n \times \min_{c_j, c_q \neq c_j} \text{dist}^2(c_j, c_q)}$	[0, +∞]	Min	$O(n^2 d \log(n))$

DB: Dataset, n: number of data objects in DB, v : center of DB, d: number of attributes, c: data objects of DB, k: number of clusters, C_i : i^{th} cluster, c_j : j^{th} member of i^{th} cluster, v_i : center of i^{th} cluster, $\text{var}(C)$: variance vector of C_p $\text{dist}(\cdot)$: distance function.

et al. [43] also worked to introduce indices based on Dunn variations and cohesion, which act well with noisy and overlapped clusters. Table I reports the definition, range, optimum value, and complexity of each of the above-discussed internal CVIs.

III. PROPOSED CVI

In this section, we examine and present some of the imperative properties of SD and propose a new internal CVI measure.

A. S-Divergence and Its Properties

Definition 1. SD presents a metric on the set of matrices A_τ of size $\tau \times \tau$ [31]. The set A_τ is a convex cone, on which SD is defined using Eq. 1.

$$D_S^2(A_\tau^i, A_\tau^j) = \log(\det(\frac{A_\tau^i + A_\tau^j}{2})) - \frac{1}{2} \log(\det(A_\tau^i A_\tau^j)) \quad (1)$$

where $\det(\cdot)$ denotes the determinant operation. D_S is a metric on the positive definite matrices (PDM) A_τ . Let ϕ_τ be a one-to-one function from $\mathfrak{R}_\tau^+ \rightarrow A_\tau$. Now examine a vector $\mathbf{t} = \{t_1, t_2, \dots, t_\tau\} \in \mathfrak{R}_\tau^+$ to generate PDM from a vector \mathbf{t} . SD is a divergence function on the cone of HPDM. A convex cone structure on the set of HPDM enables “geometric optimization”, which enables us to resolve certain problems that may be non-convex in Euclidean space but convex in manifold space, or, offers efficient optimization. Thus, the divergence function on the cone of hpd matrices has empirical and computational advantages in many applications [44].

At this juncture, we shall demonstrate that the SD meets all the necessary characteristics for becoming a distance metric, which are given below:

Proposition 1. Non-negativity: $D_S^2(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{u})) \geq 0$

Proof. The modified version of Eq. 1 is given below:

$$D_S^2(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{u})) = \log(\det(\frac{\phi_\tau(\mathbf{t}) + \phi_\tau(\mathbf{u})}{2})) + \log(\frac{1}{\sqrt{\det(\phi_\tau(\mathbf{t})\phi_\tau(\mathbf{u}))}}) \quad (2)$$

$$\Rightarrow D_S^2(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{u})) = \log(\frac{\det(\frac{\phi_\tau(\mathbf{t}) + \phi_\tau(\mathbf{u})}{2})}{\sqrt{\det(\phi_\tau(\mathbf{t})\phi_\tau(\mathbf{u}))}}) \quad (3)$$

where $\frac{\det(\frac{\phi_\tau(\mathbf{t}) + \phi_\tau(\mathbf{u})}{2})}{\sqrt{\det(\phi_\tau(\mathbf{t})\phi_\tau(\mathbf{u}))}} \geq 0$ because determinant of the PDM is always positive and numerator will be greater than or equal to denominator. $\therefore D_S^2(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{u})) \geq 0$ □

Proposition 2. Equality: $D_S^2(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{u})) = 0$ iff $\mathbf{t} = \mathbf{u}$

Proof. From proposition 1, we can write

$$D_S^2(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{u})) = \log(\frac{\det(\frac{\phi_\tau(\mathbf{t}) + \phi_\tau(\mathbf{u})}{2})}{\sqrt{\det(\phi_\tau(\mathbf{t})\phi_\tau(\mathbf{u}))}})$$

Now, if \mathbf{t} and \mathbf{u} are equal then \mathbf{u} can be replaced by \mathbf{t} in the above expression and the modified expression is

$$D_S^2(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{t})) = \log\left(\frac{\det(\frac{\phi_\tau(\mathbf{t})+\phi_\tau(\mathbf{t})}{2})}{\sqrt{\det(\phi_\tau(\mathbf{t})\phi_\tau(\mathbf{t}))}}\right) \Rightarrow \log(1) = 0$$

$$\therefore D_S^2(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{u})) = 0 \text{ iff } \mathbf{t} = \mathbf{u}.$$

Please note that we used the property that the determinant of the power of a matrix is equal to the determinant raised to that power, meaning in our case:

$$\det(\phi_\tau(\mathbf{t})\phi_\tau(\mathbf{t})) = \det(\phi_\tau(\mathbf{t}))^2$$

Proposition 3. Symmetry:

$$D_S^2(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{u})) = D_S^2(\phi_\tau(\mathbf{u}), \phi_\tau(\mathbf{t}))$$

Proof. The SD amid \mathbf{t} and \mathbf{u} is denoted as follows:

$$D_S^2(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{u})) = \log\left(\frac{\det(\frac{\phi_\tau(\mathbf{t})+\phi_\tau(\mathbf{u})}{2})}{\sqrt{\det(\phi_\tau(\mathbf{t})\phi_\tau(\mathbf{u}))}}\right) \text{ [as already noted in proposition 1]} = D_S^2(\phi_\tau(\mathbf{u}), \phi_\tau(\mathbf{t}))$$

$$\therefore D_S^2(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{u})) = D_S^2(\phi_\tau(\mathbf{u}), \phi_\tau(\mathbf{t}))$$

It implies SD also abides the symmetric metric property. \square

Proposition 4. Triangle Inequality: Suppose \mathbf{t} , \mathbf{u} , and \mathbf{z} be three vectors. Then this proposition states, the sum of the lengths of any two sides viz., $D_S(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{u}))$ and $D_S(\phi_\tau(\mathbf{u}), \phi_\tau(\mathbf{z}))$ of a triangle is greater than or equal to the length of the third side $D_S(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{z}))$. Arithmetically, $D_S(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{z})) \leq D_S(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{u})) + D_S(\phi_\tau(\mathbf{u}), \phi_\tau(\mathbf{z}))$.

Proof. Let \mathbf{t} , \mathbf{u} , and \mathbf{z} be three vectors. Then $\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{u}), \phi_\tau(\mathbf{z}) > 0$ and diagonal matrices.

$$\text{Thus } D_S^2(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{u})) = \sum_i D_S^2(t_i, u_i),$$

$$D_S^2(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{z})) = \sum_i D_S^2(t_i, z_i), \text{ and}$$

$$D_S^2(\phi_\tau(\mathbf{u}), \phi_\tau(\mathbf{z})) = \sum_i D_S^2(u_i, z_i)$$

$$\therefore D_S^2(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{z})) \leq D_S^2(\phi_\tau(\mathbf{t}), \phi_\tau(\mathbf{u})) + D_S^2(\phi_\tau(\mathbf{u}), \phi_\tau(\mathbf{z}))$$

Hence, it is showed that the SD is a metric. \square

B. Cluster Density Estimation

In this study, each cluster is modeled using a random variable characterized by a probability distribution. In practice, the underlying probability distribution of a random variable is not known in advance. Alternatively, the probability distribution of a random variable is estimated from the data objects or samples of a cluster. Therefore, each random variable is associated with a set of samples. We assume that samples are finite, independent, and identically distributed. Here, we adopt the well-known non-parametric probability estimation technique KDE to estimate the underlying distribution of the observations.

Let M be a random variable characterizing cluster C_m , where each sample, \mathbf{x} , is of d -dimensions. Then, the kernel function is obtained by multiplying the d number of Gaussian functions with bandwidth, h_l^M , where $1 \leq l \leq d$ and $d \geq 2$. Equation 4 is applied to estimate M [1], [2].

$$M(\mathbf{x}) = \frac{1}{|C_m|(2\pi)^{\frac{d}{2}} \prod_{l=1}^d h_l^M} \sum_{c_j \in C_m} \prod_{l=1}^d e^{-\frac{(x^l - c_{j,l})^2}{2h_l^M}} \quad (4)$$

where $x \in \mathcal{D}$, every cluster is defined in the same domain \mathcal{D} and we also assume that the \mathcal{D} is a bounded range of values and c_j is a j^{th} member of i^{th} cluster or $c_j \in C_i$, h_l^M is the bandwidth of the l^{th} feature and it controls the smoothing of the Gaussian kernel function. The Silverman approximation rule (Eq. 5) is considered to estimate h_l^M .

$$h_l^M = 1.06 \times \sigma_l |C_m|^{-\frac{1}{5}} \quad (5)$$

where σ_l denotes the standard deviation of C_m for the l^{th} feature.

C. S-Divergence Between Two Clusters

The SD between two clusters is stated as follows:

Definition 2. Let C_m and C_n be two clusters. The M and N are the two probability mass functions (PMFs) of C_m and C_n respectively as defined in Eq. 4 with finite or countably infinite values in a discrete domain, \mathcal{D} . The SD between C_m and C_n is computed by Eq. 6.

$$D_S^2(M, N) = \log\left(\frac{\phi_{|C_m|}(M) + \phi_{|C_m|}(N)}{2}\right) - \frac{1}{2} \log(\det(\phi_{|C_m|}(M)\phi_{|C_m|}(N))) \quad (6)$$

where we assume that M has C_m samples $M = \{x_1, x_2, \dots, x_{|C_m|}\}$ and PMF of every uncertain object is converted into diagonal matrix using $\phi_{|C_m|}(\cdot)$ function as follows:

$$\phi_{|C_m|}(M) = \begin{bmatrix} M(x_1) & 0 & \dots & 0 \\ 0 & M(x_2) & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & M(x_{|C_m|}) \end{bmatrix}$$

$$\text{and } \phi_{|C_m|}(N) = \begin{bmatrix} N(x_1) & 0 & \dots & 0 \\ 0 & N(x_2) & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & N(x_{|C_m|}) \end{bmatrix}.$$

Sometimes, it is needed to smooth a PMF of a data object thus the probability values become non-negative in a domain since SD consists of a logarithmic function as shown in Eq. 6. Thus, Eq. 7 is employed for normalizing [1].

$$N'(x) = \frac{N(x) + \beta}{1 + \beta|\mathcal{D}|} \quad (7)$$

where β is a constant and the value of β lies between an interval $[0, 1]$. The $|\mathcal{D}|$ signifies the number of possible values in \mathcal{D} . Furthermore, the sum of integral of $N'(x)$ over the entire \mathcal{D} is 1. Equation 8 is utilized to estimate error in smoothing.

$$|N'(x) - N(x)| = \left| \frac{1 - N(x)\beta}{\beta + |\mathcal{D}|} \right| \in \left[0, \frac{\max\{1, |1 - \mathcal{D}|\}}{\beta + |\mathcal{D}|}\right] \quad (8)$$

The value of β is assigned to 0.001 in this work. The $\phi_{|C_m|}$ function is used to convert probability distributions to HPDM. The HPDM are manifolds, which are similar to non-positive curvature [31]. The HPDM cone does not come with a natural similarity function for a data object, although, it has computational and empirical advantages. Now, Eq. 6 is further simplified as follows:

$$D_S^2(M, N) = \log\left(\frac{1}{2|C_m|} \times \det(\phi_{|C_m|}(M) + \phi_{|C_m|}(N))\right) + \log\left(\frac{1}{\det(\phi_{|C_m|}(M)\phi_{|C_m|}(N))}\right) =$$

$$\log\left(\frac{1}{2|C_m|} \times \frac{(M(x_1) + N(x_1))(M(x_2) + N(x_2)) \dots (M(x_{|C_m|}) + N(x_{|C_m|}))}{\sqrt{M(x_1)M(x_2) \dots M(x_{|C_m|})N(x_1)N(x_2) \dots N(x_{|C_m|})}}\right)$$

$$D_S^2(M, N) = \log\left(\frac{1}{2|C_m|} \times \frac{(M(x_1) + N(x_1))}{\sqrt{M(x_1)N(x_1)}} \times \frac{(M(x_2) + N(x_2))}{\sqrt{M(x_2)N(x_2)}} \times \dots \times \frac{(M(x_{|C_m|}) + N(x_{|C_m|}))}{\sqrt{M(x_{|C_m|})N(x_{|C_m|})}}\right)$$

$$D_S^2(M, N) = \log\left(\frac{1}{2|C_m|}\right) + \log\left(\left(\frac{M(x_1)}{N(x_1)} + \frac{N(x_1)}{M(x_1)}\right)\left(\frac{M(x_2)}{N(x_2)} + \frac{N(x_2)}{M(x_2)}\right) \dots \left(\frac{M(x_{|C_m|})}{N(x_{|C_m|})} + \frac{N(x_{|C_m|})}{M(x_{|C_m|})}\right)\right)$$

$$D_S^2(M, N) = \log\left(\frac{1}{2|C_m|}\right) + \log\left(\frac{M(x_1)}{N(x_1)}\right) + \log\left(\frac{M(x_2)}{N(x_2)}\right) + \dots + \log\left(\frac{M(x_{|C_m|})}{N(x_{|C_m|})}\right) + \log\left(1 + \frac{N(x_1)}{M(x_1)}\right)$$

$$+ \log\left(1 + \frac{N(x_2)}{M(x_2)}\right) + \dots + \log\left(1 + \frac{N(x_{|C_m|})}{M(x_{|C_m|})}\right)$$

$$D_S^2(M, N) = \log\left(\frac{1}{2|C_m|}\right) + \sum_{x \in \mathcal{D}} \log\left(\sqrt{\frac{M(x)}{N(x)}}\right) \left(1 + \frac{N(x)}{M(x)}\right)$$

Finally, the SD between M and N is expressed as follows:

$$D_S^2(M, N) = \log\left(\frac{1}{2|C_m|}\right) + \sum_{x \in \mathcal{D}} \log\left(\frac{|C_m| \prod_{l=1}^d h_l^M \sum_{c_j \in C_m} \prod_{l=1}^d e^{-\frac{(x^l - c_{j,l})^2}{2h_l^M}}}{|C_m| \prod_{l=1}^d h_l^M \sum_{c_j \in C_n} \prod_{l=1}^d e^{-\frac{(x^l - c_{j,l})^2}{2h_l^M}}}\right) \left(1 + \frac{|C_m| \prod_{l=1}^d h_l^M \sum_{c_j \in C_n} \prod_{l=1}^d e^{-\frac{(x^l - c_{j,l})^2}{2h_l^M}}}{|C_n| \prod_{l=1}^d h_l^N \sum_{c_j \in C_n} \prod_{l=1}^d e^{-\frac{(x^l - c_{j,l})^2}{2h_l^N}}}\right)$$

D. The Proposed Internal CVI

The proposed internal CVI (PM) is based on CM and SM. So, the values of CM and SM need to be computed before calculating PM. The CM indicates the closeness or similarity of data objects in a cluster. Moreover, it is an average CM of all k clusters. The CM of every cluster, C_p , is calculated by Eq. 9. It is an average of aggregated squared S-distance (SD) of a cluster data object c_j to its center v_r .

$$CM = \frac{1}{k} \sum_{i=1}^k \frac{1}{|C_i|} \sum_{c_j \in C_i} D_{SD}(c_j, v_i) \quad (9)$$

where D_{SD} is the SD, which can be defined mathematically using Eq. 10.

Definition 3. define $D_{SD}: \mathfrak{R}_+^d \times \mathfrak{R}_+^d \rightarrow \mathfrak{R}_+ \cup \{0\}$ as

$$D_{SD}(c_j, v_i) = \sum_{l=1}^d [\log((c_{j,l} + v_{i,l})/2) - (\log(c_{j,l}) + \log(v_{i,l}))/2] \quad (10)$$

Equation 10 shows a point-to-point distance measure labeled as the SD that is motivated by the SD. It is defined in the open cone of PDM. Moreover, Eq. 10 shows that if two data objects with the same Euclidean distance are close to the origin, then data objects will have a larger SD compared to when they are far from the origin. This property can be applied to find the properties of clusters with varying sizes and densities. Furthermore, SD is neither an f-divergence nor a Bregman divergence and is invariant under the Hadamard product [45].

The CM ranges from 0 to ∞ , where a low value is appropriate for a clustering configuration. The SM determines the magnitude of separation between clusters. The SD-based SM is calculated in this study by Eq. 11.

$$SM = \frac{2}{k(k-2)} \sum_{i=1}^k \sum_{j=i+1}^k D_S(M_i, M_j) \quad (11)$$

where M_i and M_j are the PMFs of clusters C_i and C_j respectively. The SM lies in the interval $[0, \infty)$, where a high value implies good clustering. The PM is a ratio of the CM to the proposed SM, and it is estimated using Eq. 12.

$$PM = \frac{CM}{SM} \quad (12)$$

Good clustering is characterized by a low CM and a high SM of clusters. Therefore, a smaller value of PM is suitable for a clustering configuration. Sometimes, it is required to normalize the SM, and thus its value becomes non-zero in a domain since the zero value of SM will make an undefined value of the proposed index, PM, as shown in Eq. 12. Hence, Eq. 11 is further normalized.

$$SM = \frac{2}{k(k-2)} \sum_{i=1}^k \sum_{j=i+1}^k D_S(M_i, M_j) + \frac{\delta}{k^2} \quad (13)$$

where δ is a constant and the value of $\delta \rightarrow 0$, further estimated error in normalization is $\frac{\delta}{k^2}$ which is less significant in the possible range of SM. Normalized SM will be used throughout the paper to avoid an undefined value.

E. Complexity Analysis

The complexity associated with CM and SM is $O(nd)$ and $O(k^2dE)$ respectively, where E is the number of steps to estimate the SD between two clusters. The complexity of PM is represented by $O(nd + k^2dE)$ since $n \geq d$ and $n \geq k$ is considered in this study. Thus the complexity of the proposed CVI is linear.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A laptop Intel(R) Core(TM) i7-2620M CPU@2.70GHz and 4-GB RAM running on Windows 10 having a 64-bits Python 3.6.5 compiler are considered for this study. All the work is carried out in Spyder 3.2.8's Python development environment.

A. Description of Databases

A total of 10 databases of two classes, namely synthetic and real-world are considered in this work to prove the effectiveness of the PM over some of the most popular existing internal CVIs. **Synthetic databases:** Three databases, namely Blobs, Varied Distributed data, and Anisotropically Distributed Data, are created in this study. The title of the databases, the total number of data objects in each database, the total number of features in each data object, and the number of clusters are noted in Table II. The Blobs database is produced by an isotropic Gaussian function with three classes having 1500 data objects or samples and two features. The varied distributed data is produced with varied variance in the data and has 1500 samples with 3 classes in 2D space, whereas Anisotropically distributed database is generated by transforming the data, which is Anisotropically distributed or aligned on a specific axis. This database also has 1500 samples, three classes, and two features. **UCI and Kaggle repository databases:** Seven popular realistic databases, viz., Digits, Iris, Wine, Avila, Shuttle, Breast Cancer, and Letter Recognition, are adopted from the UCI repository [46], [47]. The short description of each of these UCI databases is also reported in Table II. All the databases are renamed as DB_i , where i varies from 1 to 10.

TABLE II. DATASETS CHARACTERISTICS

S. No.	Datasets	No of data objects	No of features	Clusters
1	Varied distributed data (DB1)	1500	2	3
2	Anisotropically distributed data (DB2)	1500	2	3
3	Blobs (DB3)	1500	2	3
4	Breast cancer database (DB4)	569	30	2
5	Iris database (DB5)	150	4	3
6	Wine database (DB6)	178	13	3
7	Avila database (DB7)	10430	10	12
8	Digits database (DB8)	1797	64	10
9	Letter recognition database (DB9)	20000	16	26
10	Shuttle database (DB10)	43500	9	7

B. Results and Comparison

A couple of experiments are conducted to prove the effectiveness of PM over some of the existing internal CVIs in different scenarios, which are as follows:

1. The Impact of Monotonicity

The first experiment aims to study the monotonicity behavior of three internal CVIs, namely RMSSTDI, RSI, and MHI. Three synthetic databases, namely DB_1 , DB_2 , and DB_3 are considered, where clusters are well-separated. Fig. 2 (a), (c), and (e) plot the datasets DB_1 , DB_2 , and DB_3 along the x and y axes on a 2D plane, respectively. Here, fuzzy k-means (FKM) is applied to the three databases mentioned above, and the values of RMSSTDI, RSI, and MHI are computed, which are labeled as FKM-RMSSTDI, FKM-RSI, and FKM-MHI, respectively. Fig. 2 (b), (d), and (f) show the values of FKM-RMSSTDI, FKM-RSI, and FKM-MHI, respectively, that are obtained by varying the number of

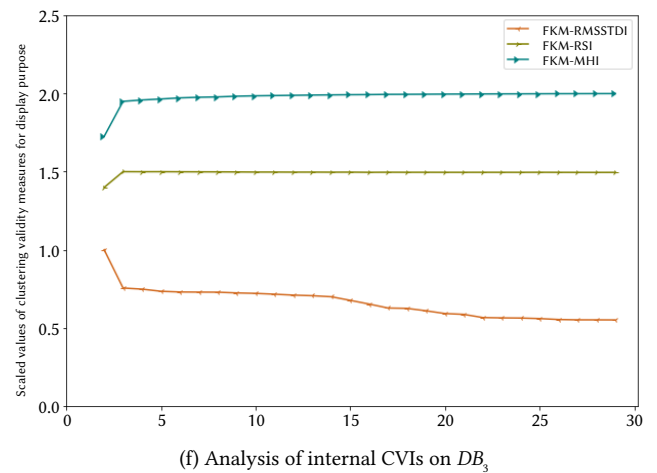
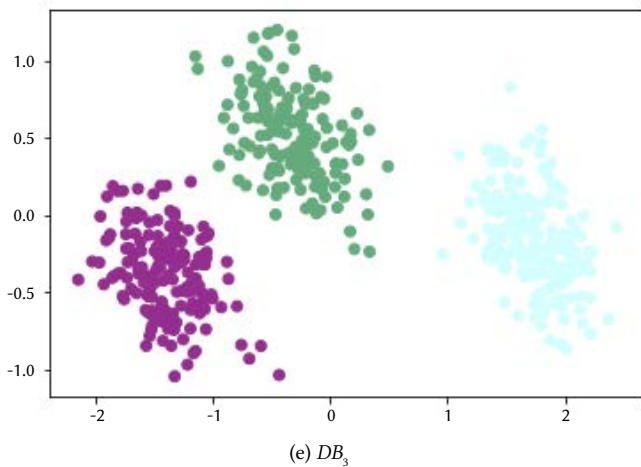
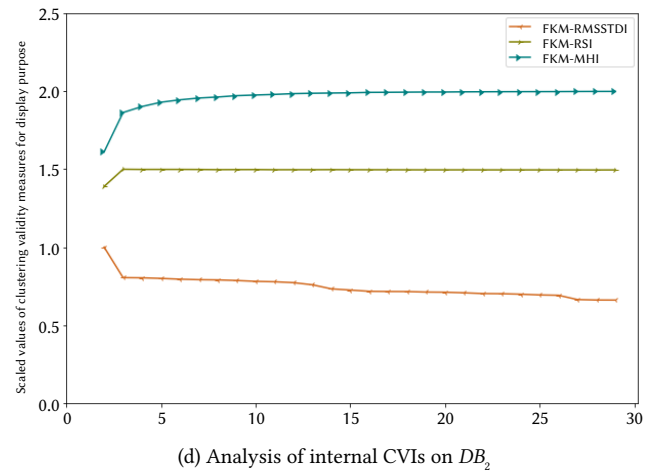
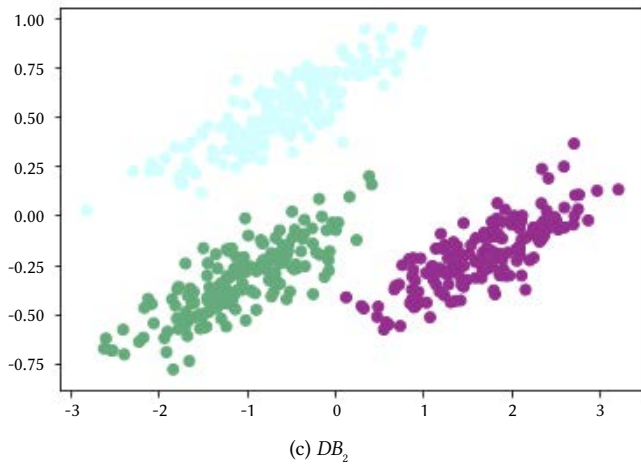
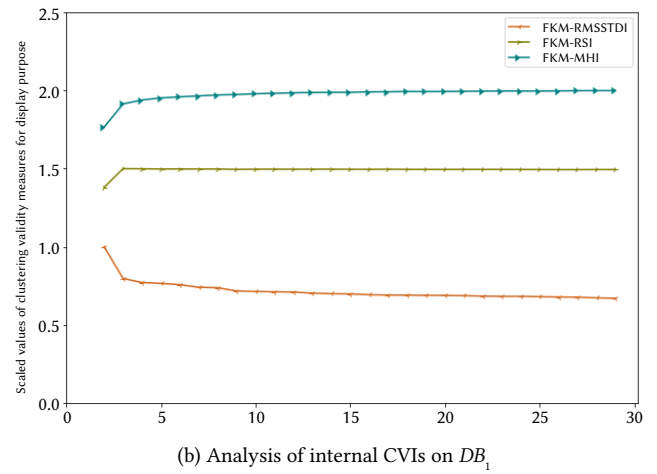
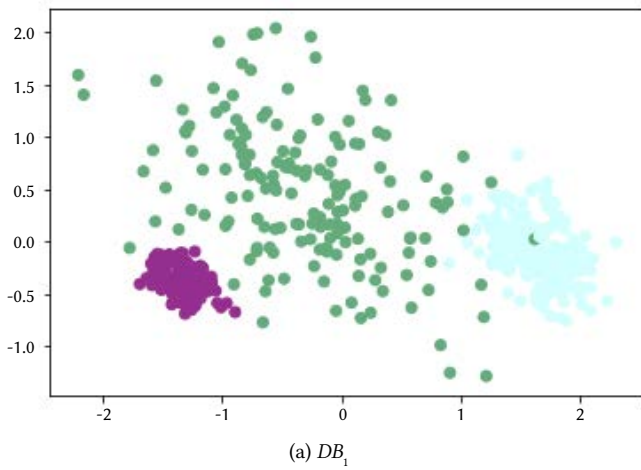


Fig. 2. DB_1 , DB_2 , and DB_3 are plotted on the plane, different classes are shown with different colors and result of internal CVIs on database in the right.

clusters, k , from 2 to 29 as inputs because the datasets discussed in Table II have an actual number of clusters in the range of 2 to 26. The other information on the results is not pertinent to this experiment. The vertical axis of curves or graphs in Fig. 2 is scaled for better visualization or analysis. When the value of k increases then value of numerator in $RMSSTDI = \frac{\sum_{i=1}^k \sum_{c_j \in C_i} \|c_j - v_i\|^2}{d(n-k)}$ will decrease. The value of $(n - k)$ is regarded as a constant because $k \ll n$. Therefore, RMSSTDI decreases with an increase in the k -value in Fig. 2 (b), (d), and (f). Further, RSI specifies a ratio of between clusters sum of squares to the total sum of squares. Hence, RSI increases as the value of k increases, as shown in Fig. 2 (b), (d), and (f). Similarly, MHI increases as the value of k

increases, according to Fig. 2 (b), (d), and (f), because with an increase in k more pairs of distances are calculated. Furthermore, RMSSTDI is only based on CM, and RSI and MHI rely only on SM. According to the property of monotonicity, the curves of RMSSTDI, RSI, and MHI will be either downward or upward. It is quoted that the value of k is optimal at the "elbow" point, where a shift in the curve appears. Thus, the empirical results in Fig. 2 prove that the RMSSTDI, RSI, and MHI monotonically decrease or increase as the number of clusters, k , increases in the range from 2 to 29. However, the determination of a shift in the curve is rather a tedious and subjective task, thus the monotonicity is not discussed in the further sections.

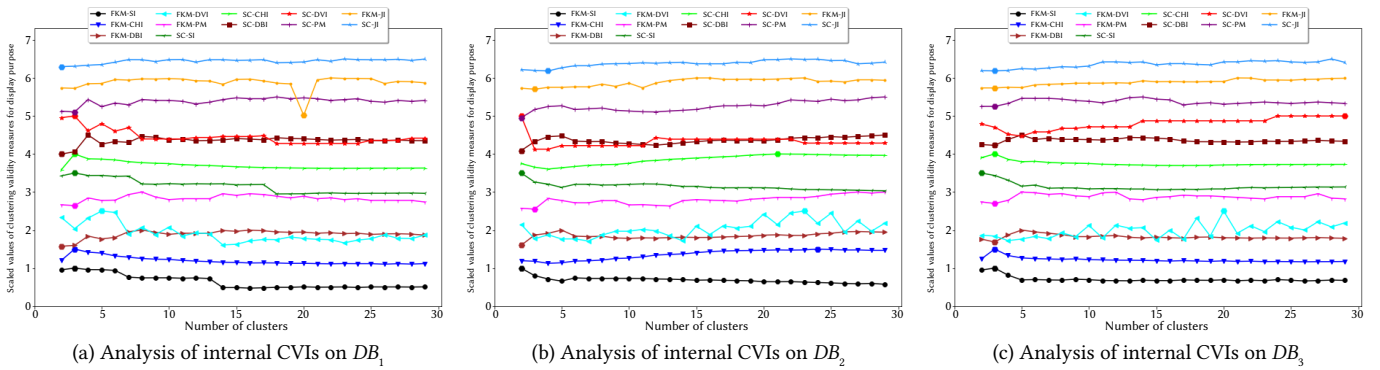


Fig. 3. An analysis of internal CVIs on well-separated databases.

2. The Impact of Well-Separated Clusters

The aim of the 2nd experiment is to determine the optimal value of k for the databases, where well-separated clusters are present. The steps involved in estimating the optimal value of k for the best partitions using internal CVIs are as follows:

- Step 1: Initialize a clustering algorithm before applying it to a database.
- Step 2: A set of parameters of the algorithm is fixed in order to achieve clustering results.
- Step 3: Calculate the corresponding internal CVIs after clustering.
- Step 4: Select the optimal value of internal CVIs for best partitions.

Here, the values of six internal CVIs viz., SI, CHI, DBI, DVI, JI, and PM are computed after applying FKM and spectral clustering (SC) [48] on three databases, namely DB_1 , DB_2 , and DB_3 and results are reported in Fig. 3 (a), (b), and (c) respectively. The FKM-SI, FKM-CHI, FKM-DBI, FKM-DVI, FKM-JI, and FKM-PM specify the values of SI, CHI, DBI, DVI, JI, and PM after executing FKM while SC-SI, SC-CHI, SC-DBI, SC-DVI, SC-JI, and SC-PM are employed to represent the values of SI, CHI, DBI, DVI, JI, and PM after applying SC. Fig. 3 displays the values of FKM-SI, FKM-CHI, FKM-DBI, FKM-DVI, FKM-JI, FKM-PM, SC-SI, SC-CHI, SC-DBI, SC-DVI, SC-JI, and SC-PM that are obtained by varying the value of k in the range of 2 to 29. The optimal values of CVIs labeled by a hexagon marker in Fig. 3 specify either maximum or minimum values, which demonstrate the actual values of k in the databases. It is clear from Fig. 3 (a) that SC-PM, FKM-PM, FKM-JI, SC-JI, SC-DVI, SC-CHI, SC-SI, FKM-CHI, and FKM-SI determine the optimal value of k , which is the same as the exact number of clusters in DB_1 . Moreover, the remaining CVIs produce values of k , which are closer to the actual number of clusters. It is also observed from Fig. 3 (b) that the FKM-PM and FKM-JI compute the optimal number of clusters, which are equal to the real number of clusters in DB_2 . Furthermore, FKM-SI, FKM-DBI, SC-SI, SC-DBI, SC-DVI, SC-PM, FKM-JI, and SC-JI are also in proximity to the optimal clusters. On the other hand, the remaining CVIs are not near-optimal results. Fig. 3 (c) shows the results of DB_3 and that FKM-SI, FKM-CHI, FKM-DBI, FKM-PM, SC-DBI, SC-PM, FKM-JI, SC-JI, and SC-CHI achieve the optimal value for the clusters.

3. The Impact of Slightly Overlapped Clusters

The third experiment aims to decide the optimal value of k for the databases, namely DB_4 , DB_5 , and DB_6 , where slightly overlapping clusters are present. However, principal component analysis is adopted in exploratory data analysis by transforming the data to a new coordinate system in the case of high-dimensional data and then plotting the first two principal components [49], [50]. The first two principal components of datasets DB_4 , DB_5 , and DB_6 are mapped on a 2D plane, which are displayed in Fig. 4 (a), (c), and (e), respectively. Here, slightly overlapping clusters are denoted by different colors.

Again, the values of six internal CVIs, viz. SI, CHI, DBI, DVI, JI, and PM, are computed after applying FKM and SC on the three databases mentioned above, and the outcomes are noted in Fig. 4 (b), (d), and (f), respectively. Here, we run the clustering algorithms for different values of k in the range of 2 to 29. We can find out the exact values of k by considering the optimum values of the curves of the FKM-PM and SC-PM in most cases. Moreover, PM always helps to decide the exact value of k because of the use of non-linear similarity measures.

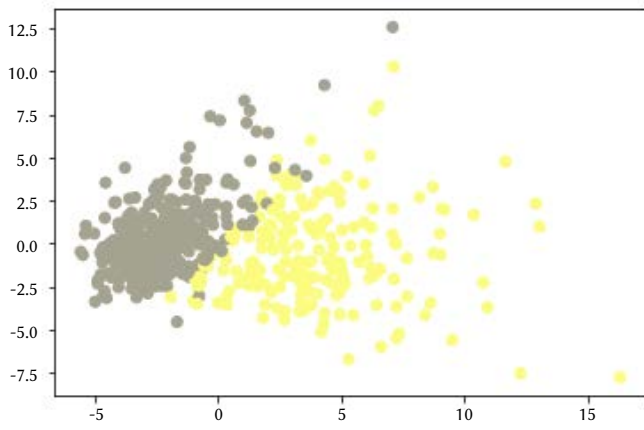
4. The Impact of Highly Overlapped Clusters

The focus of the fourth experiment is to estimate the optimal value of k for the databases, namely DB_7 , DB_8 , DB_9 , and DB_{10} , where clusters are highly significant. The first two principal components of datasets DB_7 , DB_8 , DB_9 , and DB_{10} are mapped on a 2D plane, which are displayed in Fig. 5 (a), (c), (e), and (g), respectively. Here, different colors are employed to represent clusters. Again, the values of six internal CVIs, viz. SI, CHI, DBI, DVI, JI, and PM, are calculated after applying FKM and SC on the four databases stated above, and the results are displayed in Fig. 5 (b), (d), (f), and (h), respectively. Here, both the clustering algorithms execute for different values of k in the range of 2 to 29. Focusing on the results, PM determines the optimal k for DB_7 and DB_8 . But FKM-DBI, FKM-DVI, SC-SI, and SC-PM compute a value close to it. Furthermore, FKM-PM, SC-PM, and SC-DVI find the optimal k for DB_9 , and FKM-DBI and FKM-DVI are not far from them. Finally, for DB_{10} , SC-PM and SC-DVI find the optimal k and FKM-DBI, FKM-DVI, FKM-PM, SC-DBI, and SC-JI compute a close value.

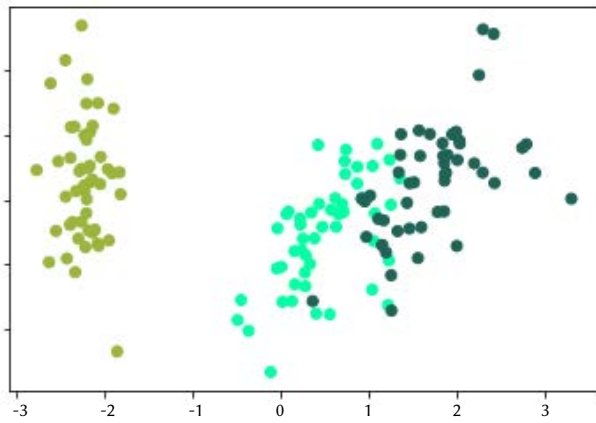
5. The Impact of Noise

The purpose of the 5th experiment is to determine how robust the proposed internal CVI named PM is against noisy features. First, noisy facets are included in the three well-separated databases, namely DB_1 , DB_2 , and DB_3 . Here, a noisy feature is produced by considering uniform random distribution in the limit of the length and size similar to features of the original database. The number of features will be doubled in a database after adding noisy features. The impact of noisy features is then analyzed in this study. Databases are shown in Fig. 6 (a), (c), and (e). Again, the values of six internal CVIs, viz. SI, CHI, DBI, DVI, JI, and PM, are estimated after applying FKM and SC to the three noisy databases presented above, and the results are portrayed in Fig. 6 (b), (d), and (f), respectively. Here, both the clustering algorithms execute for different values of k in the range of 2 to 29. It is clear from Fig. 6 that DBI and DVI are affected by noise and face difficulty while determining the optimum value of k . Further, the curve of CHI is close to the optimal number of clusters in the case of a noisy DB_2 database. On the other hand, the optimum values of SI, JI, and PM are closer to the exact values of k .

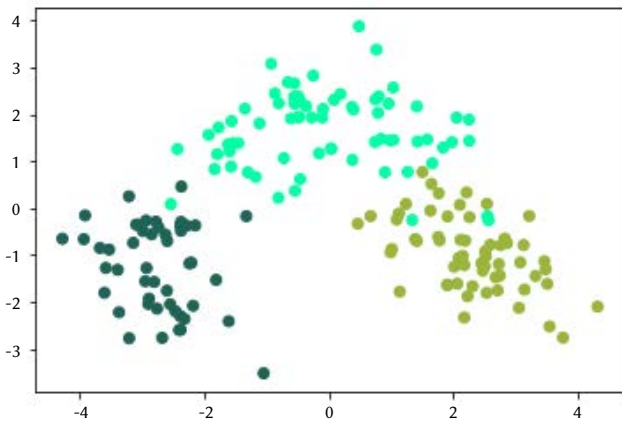
We can conclude from the five experiments conducted above that the proposed internal CVI named PM successfully ascertains the optimal number of clusters for most databases. On the other hand,



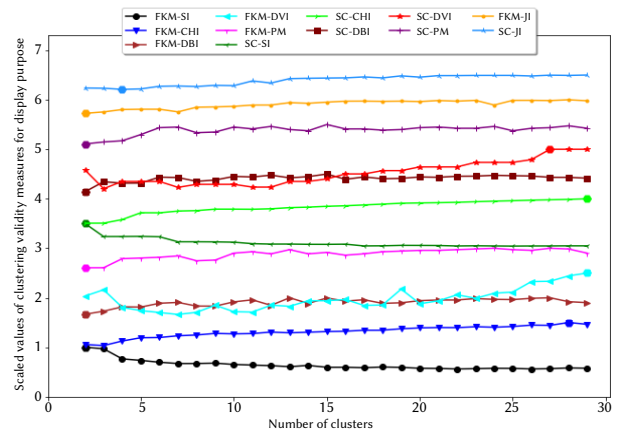
(a) DB_4



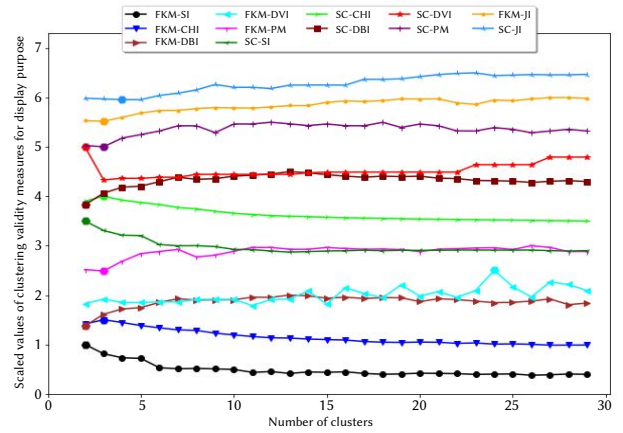
(c) DB_5



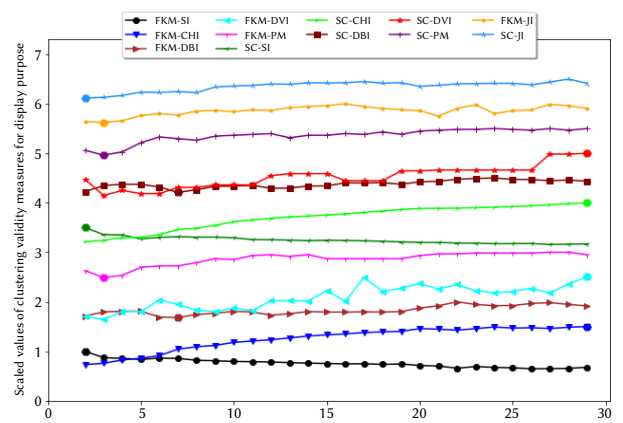
(e) DB_6



(b) Analysis of internal CVIs on DB_4



(d) Analysis of internal CVIs on DB_5



(f) Analysis of internal CVIs on DB_6

Fig. 4. In the left first two principal components of the DB_4 , DB_5 , and DB_6 are plotted on the plane, to display the first and second corresponding vectors of the data matrix along the axes, different classes are shown with different colors and result of internal CVIs on database in the right.

JI, SI, CHI, DBI, and DVI face difficulty while estimating the exact number of clusters due to various degrees of overlapping between clusters and noise in the databases.

6. The Comparative Analysis

Finally, the PM is compared with five popular internal CVIs, namely SI, CHI, DBI, DVI, and JI, after applying four clustering algorithms, viz. FKM, SC, Density-based Spatial Clustering of Applications with Noise (DBSCAN), and Density Peak Clustering (DPC) [51] on the ten databases mentioned in Section IV A. Here, FKM and SC take the exact number of clusters as inputs, whereas DBSCAN and DPC compute

the number of clusters automatically. The values of six internal CVIs, including the PM, are reported in Table III. The mean (μ) and standard deviation (σ) obtained by the four clustering algorithms of each CVI are also noted in the last column of Table III. The μ and σ of the PM are highlighted by bold characters. A smaller value of σ in percentage specifies well-separated and compact clusters. In other words, a smaller value of σ demonstrates that the clustering configuration is appropriate. It is clear from Table III that the PM consistently outperforms five considered internal CVIs on ten databases in different scenarios presented in Table IV. Therefore, PM can be a great choice while evaluating clustering results.

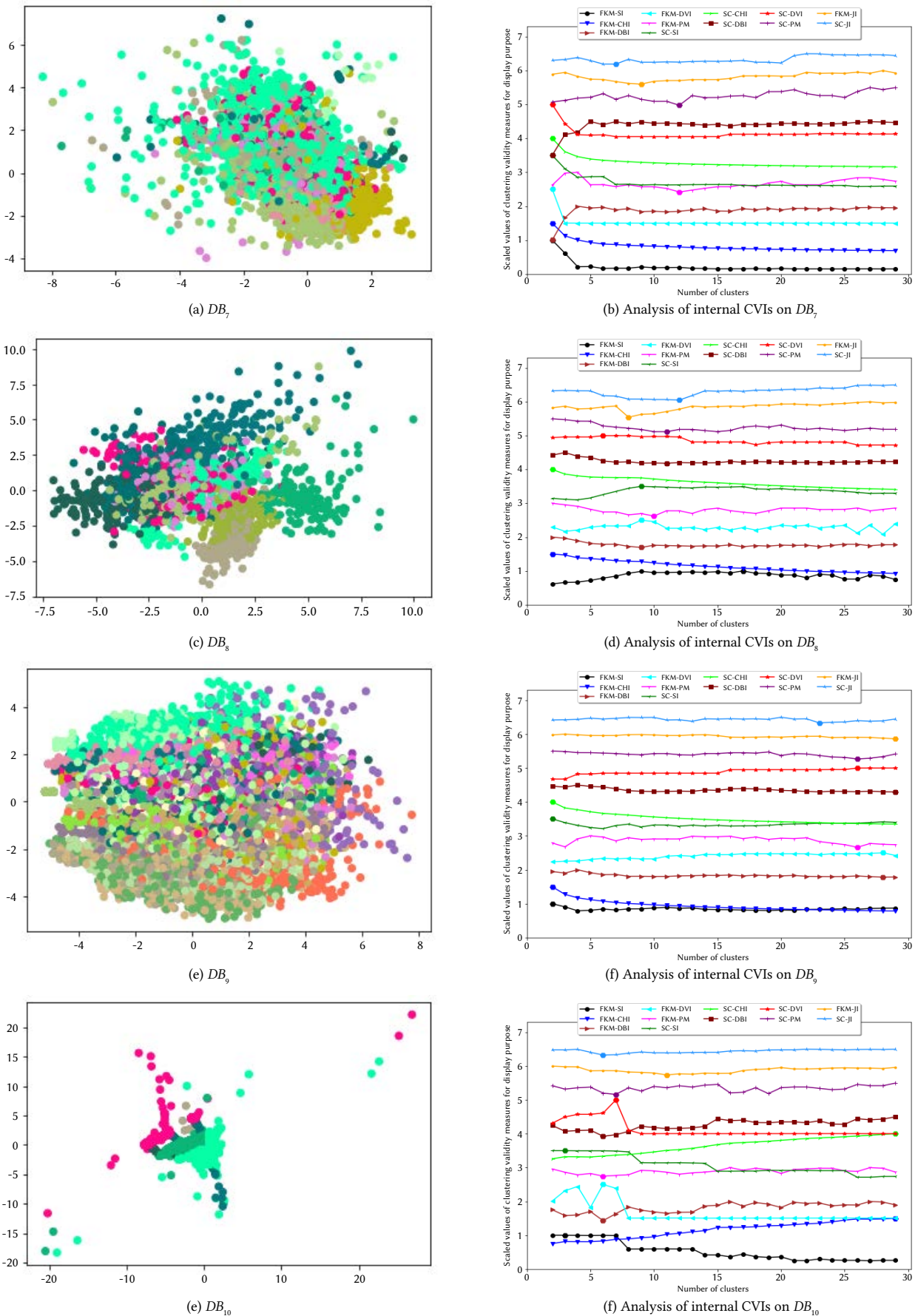
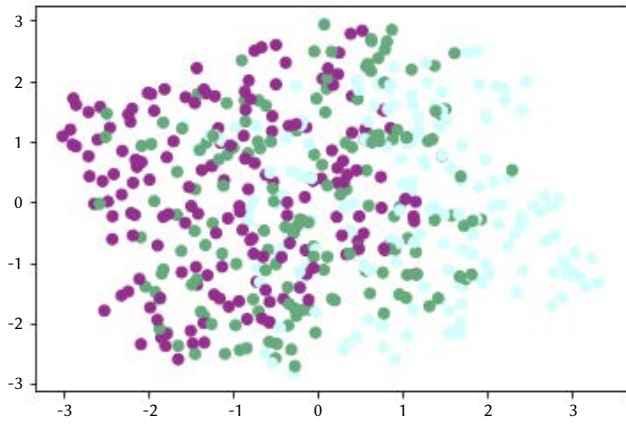


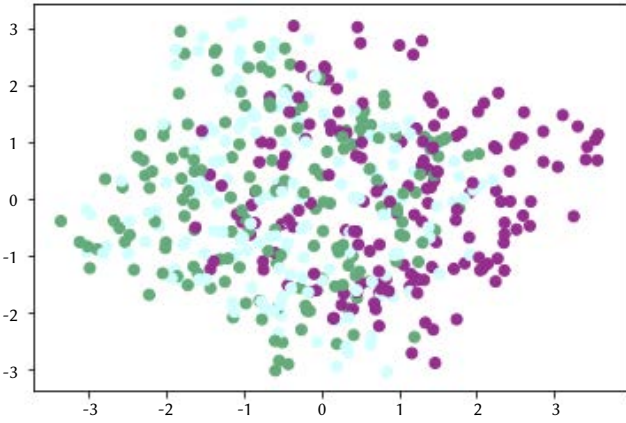
Fig. 5. In the left first two principal components of the DB_7 , DB_8 , DB_9 , and DB_{10} are plotted on the plane, to display the first and second corresponding vectors of the data matrix along the axes, different classes are shown with different colors and result of internal CVIs on database in the right.

TABLE III. COMPARATIVE ANALYSIS OF INTERNAL CVIs USING CLUSTERING ALGORITHMS

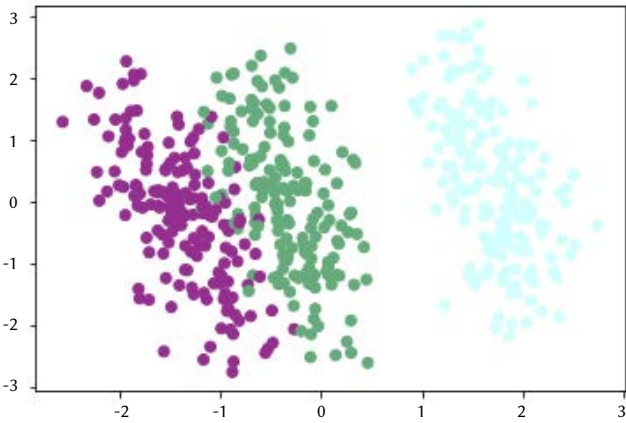
Dataset	CVI	FKM	SC	DBSCAN	DPC	$\mu \pm \sigma \%$
DB_1	SI	0.6468	0.61745	0.57755	0.62765	0.61736 \pm 8.79207
	CHI	5451.4914	3810.65643	3792.7863	4756.62157	4452.88893 \pm 18.04829
	DBI	0.56956	0.54442	0.8377	0.61122	0.64073 \pm 20.94112
	DVI	0.00785	0.01372	0.02252	0.03266	0.01919 \pm 56.37716
	JI	37.00884	39.56956	45.42931	34.47878	39.12162 \pm 11.98999
	PM	29.56956	30.33441	34.71643	28.57965	30.80001 \pm 4.72936
DB_2	SI	0.63551	0.48386	0.40725	0.50914	0.50894 \pm 18.63688
	CHI	3883.88156	3302.55351	5465.96704	3803.73873	4114.03521 \pm 22.78248
	DBI	0.49246	0.68642	0.71943	0.71491	0.65331 \pm 16.56517
	DVI	0.00899	0.0069	0.00897	0.00376	0.00716 \pm 34.47394
	JI	44.00376	41.71511	57.67286	43.40933	46.70027 \pm 15.80087
	PM	23.48942	30.68531	28.71825	22.72414	26.40428 \pm 14.78514
DB_3	SI	0.4863	0.42646	0.33207	0.46153	0.42659 \pm 15.85287
	CHI	2011.98126	1601.38907	1413.97578	1481.83841	1627.29613 \pm 16.46313
	DBI	0.73157	0.78999	0.84494	0.82526	0.79794 \pm 6.23412
	DVI	0.00825	0.01911	0.01358	0.00881	0.01244 \pm 40.60672
	JI	47.79319	43.01848	49.44894	44.48351	46.18603 \pm 6.39381
	PM	35.83244	39.79319	40.84494	37.81437	38.57124 \pm 5.74615
DB_4	SI	0.69726	0.50825	0.509	0.67526	0.59744 \pm 17.23188
	CHI	1300.20823	1089.92944	1245.56763	1251.53446	1221.80994 \pm 8.52372
	DBI	0.5044	0.62932	0.60906	0.55185	0.57366 \pm 9.87326
	DVI	0.01731	0.00726	0.01246	0.02148	0.01463 \pm 41.98165
	JI	60.5044	68.01731	74.07588	63.92288	66.6 \pm 8.76054
	PM	43.51121	49.63143	48.60891	41.56075	45.82808 \pm 7.46948
DB_5	SI	0.55282	0.55432	0.68674	0.68105	0.61873 \pm 12.16705
	CHI	561.62776	558.05804	502.82156	513.92455	534.10798 \pm 8.69226
	DBI	0.66197	0.64325	0.37927	0.39431	0.51970 \pm 29.59093
	DVI	0.09881	0.12181	0.338	0.07651	0.15901 \pm 76.31654
	JI	31.65626	32.11279	43.19802	32.38334	34.83760 \pm 16.02200
	PM	12.6709	14.65626	15.38275	13.40429	14.02855 \pm 5.63466
DB_6	SI	0.56448	0.57114	0.56067	0.56203	0.56458 \pm 6.23471
	CHI	552.85171	561.81566	670.62599	708.08668	623.34501 \pm 12.48562
	DBI	0.53573	0.53424	0.55357	0.54434	0.54197 \pm 1.64643
	DVI	0.02237	0.01626	0.0374	0.03399	0.02751 \pm 35.91714
	JI	48.01626	58.53573	51.64375	49.6353	51.95776 \pm 8.91017
	PM	27.53573	30.53424	31.55357	31.49413	30.27942 \pm 0.82341
DB_7	SI	0.1937	0.12995	0.1385	0.11951	0.14542 \pm 22.77166
	CHI	5285.5617	4519.20875	4333.76871	4212.35646	4587.72391 \pm 10.50701
	DBI	1.12112	1.29988	1.01121	0.8937	1.08148 \pm 15.96829
	DVI	0.00182	0.00529	0.00197	0.00194	0.00276 \pm 61.38810
	JI	24.12043	28.12995	35.10793	27.69293	28.76281 \pm 15.97742
	PM	8.12138	8.28694	7.57481	6.22372	7.55171 \pm 12.39664
DB_8	SI		0.1785	0.18289	0.17863	0.18066 \pm 9.87675
	CHI	169.36261	161.20475	162.1034	171.6	166.07133 \pm 3.12864
	DBI	1.9	1.88899	1.89937	1.84913	1.89023 \pm 1.63817
	DVI	0.21933	0.26126	0.17384	0.19023	0.21117 \pm 18.15176
	JI	42.87789	39.26069	49.92082	41.99865	43.51451 \pm 10.43365
	PM	15.92192	18.79859	18.90038	15.83872	17.36490 \pm 1.34074
DB_9	SI	0.1463	0.152	0.14713	0.139	0.14630 \pm 6.85984
	CHI	142	496	146	7167	1376.25764 \pm 6.35297
	DBI	1.6855	1.63312	1.64295	1.35005	1.57791 \pm 9.73410
	DVI	0.04536	0.04307	0.04136	0.04036	0.04254 \pm 5.14657
	JI	82.65005	96.04536	99.02207	94.98688	93.17609 \pm 7.75123
	PM	59.6855	66.62	60.63995	56.65005	60.89963 \pm 3.57705
DB_{10}	SI	0.97878	0.96967	0.97987	0.58508	0.87835 \pm 22.26525
	CHI	15723.30982	14946.92039	12879.555	16331.2233	14970.25213 \pm 10.05018
	DBI	0.34179	0.25082	0.3709	0.44054	0.35101 \pm 22.39261
	DVI	0.13045	0.24059	0.04701	0.09064	0.12717 \pm 65.21501
	JI	61.68367	58.31793	64.39526	63.44635	61.96080 \pm 4.31864
	PM	40.33581	39.25111	43.36991	41.43915	41.09900 \pm 4.27706



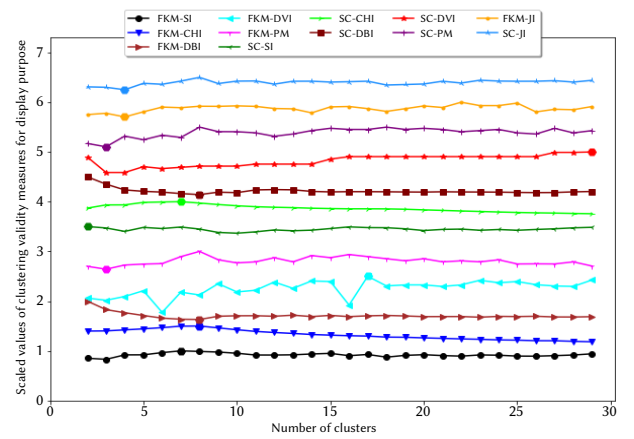
(a) Noisy- DB_1



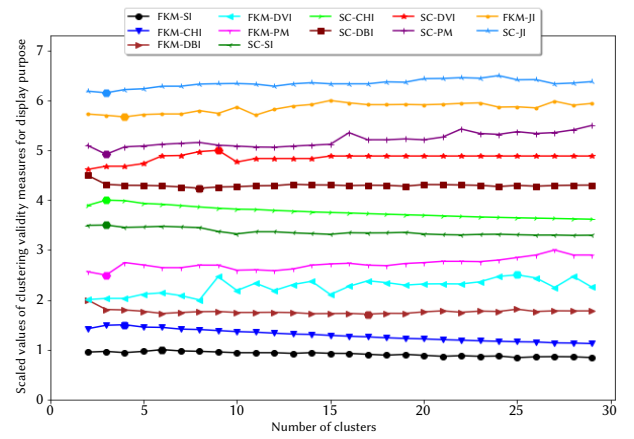
(c) Noisy- DB_2



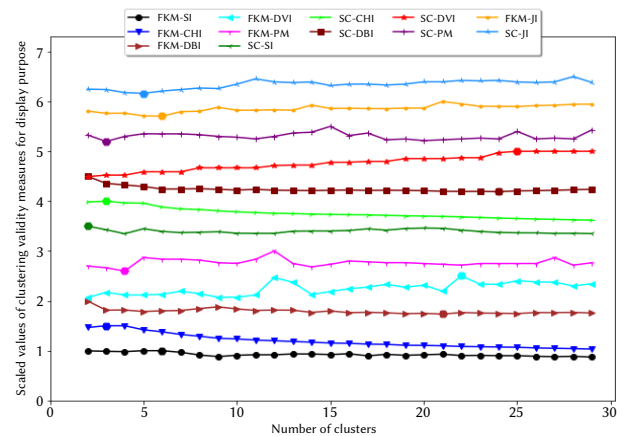
(e) Noisy- DB_3



(b) Analysis of internal CVIs on Noisy- DB_1



(d) Analysis of internal CVIs on Noisy- DB_2



(f) Analysis of internal CVIs on Noisy- DB_3

Fig. 6. In the left noisy-databases are plotted on the plane, different classes are shown with different colors and result of internal CVIs on noisy-database in the right.

TABLE IV. THE OVERALL REVIEW OF SOME INTERNAL CVIS

Index	Well-separated	Slightly-separated	Highly-overlapped	Noise
SI	G	G	X	A
CHI	G	G	X	A
DVI	G	A	X	X
DBI	A	G	X	A
JI	G	G	A	A
PM	G	G	G	G

V. CONCLUSION

Internal CVIs are employed frequently in clustering to measure the goodness of the clustering algorithms without taking any external inputs. Most of the existing internal CVIs depend on CM and the geometric distance-based SM when computing the distance between cluster centers. The previous studies showed that such CVIs are not capable of producing accurate results, especially when the clusters of a database are highly overlapping. As a remedy, we introduce a new internal CVI, PM, using a modified CM and an updated SM based on the notion of SD. Moreover, SD is defined on the cone of HPDM and is

shown to have experimental and computational advantages over the other approaches in many applications. On the other hand, SD is a point-to-point distance measure that is motivated by the definition of SD. It is defined in the open cone of PDM. Initially, clusters of a database are modeled using density functions by applying a non-parametric kernel density estimation method. The PM is defined as the ratio of the modified CM to the updated SM. A smaller value of the PM indicates that the clustering configuration is appropriate. Empirical results illustrate that the PM is proficient in determining the exact number of clusters and the best partition for several superficial and realistic databases, including the database with arbitrary cluster shapes. The proposed internal CVI faces difficulty in ascertaining the optimal number of clusters when noisy features are included in a few databases. In addition, the proposed internal CVI works efficiently for databases having only numerical attributes. The latter two aspects deserve further study. In future work, SD may be explored to develop an external CVI.

ACKNOWLEDGMENT

This work is partially supported by the project “Smart Solutions in Ubiquitous Computing Environments”, Grant Agency of Excellence (under ID: UHKFIM-GE-2023), University of Hradec Kralove, Faculty of Informatics and Management, Czech Republic.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- [1] K. K. Sharma, A. Seal, “Modeling uncertain data using monte carlo integration method for clustering,” *Expert Systems with Applications*, vol. 137, pp. 100-116, 2019.
- [2] K. K. Sharma, A. Seal, “Clustering analysis using an adaptive fused distance,” *Engineering Applications of Artificial Intelligence*, vol. 96, p. 103928, 2020.
- [3] A. Seal, A. Karlekar, O. Krejcar, E. Herrera-Viedma, “Performance and convergence analysis of modified c-means using jeffreys-divergence for clustering,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 2, pp. 141-149, 2021.
- [4] M. Martín Merino, A. J. López Rivero, V. Alonso, M. Vallejo, A. Ferreras, “A clustering algorithm based on an ensemble of dissimilarities: An application in the bioinformatics domain,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 6, pp. 6-13, 2022.
- [5] E. Asensio, A. Almeida, A. Galiano, J.-M. Martín- Álvarez, “Using customer knowledge surveys to explain sales of postgraduate programs: A machine learning approach,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 3, pp. 96-102, 2022.
- [6] F. A. Ozbay, B. Alatas, “Fake news detection within online social media using supervised artificial intelligence algorithms,” *Physica A: Statistical Mechanics and its Applications*, vol. 540, p. 123174, 2020.
- [7] B. K. Dedetürk, B. Akay, “Spam filtering using a logistic regression model trained by an artificial bee colony algorithm,” *Applied Soft Computing*, vol. 91, p. 106229, 2020.
- [8] S. Munusamy, P. Murugesan, “Modified dynamic fuzzy c-means clustering algorithm—application in dynamic customer segmentation,” *Applied Intelligence*, pp. 1–21, 2020.
- [9] I.-C. Wu, H.-K. Yu, “Sequential analysis and clustering to investigate users’ online shopping behaviors based on need-states,” *Information Processing & Management*, vol. 57, no. 6, p. 102323, 2020.
- [10] A. Sivanathan, H. H. Gharakheili, V. Sivaraman, “Detecting behavioral change of iot devices using clustering-based network traffic modeling,” *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7295–7309, 2020.
- [11] A. Das, J. Nayak, B. Naik, U. Ghosh, “Generation of overlapping clusters constructing suitable graph for crime report analysis,” *Future Generation Computer Systems*, vol. 118, pp. 339–357, 2021.
- [12] A. K. Tripathi, K. Sharma, M. Bala, A. Kumar, V. G. Menon, A. K. Bashir, “A parallel military-dog-based algorithm for clustering big data in cognitive industrial internet of things,” *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 2134–2142, 2021, doi: 10.1109/TII.2020.2995680.
- [13] M. Landauer, F. Skopik, M. Wurzenberger, A. Rauber, “System log clustering approaches for cyber security applications: A survey,” *Computers & Security*, vol. 92, p. 101739, 2020.
- [14] A. K. Abasi, A. T. Khader, M. A. Al-Betar, S. Naim, S. N. Makhadmeh, Z. A. A. Alyasseri, “Link-based multi-verse optimizer for text documents clustering,” *Applied Soft Computing*, vol. 87, p. 106002, 2020.
- [15] S. Lin, K. Schorpp, I. Rothenaigner, K. Hadian, “Image-based high-content screening in drug discovery,” *Drug discovery today*, 2020.
- [16] A. Belhadi, Y. Djenouri, J. C.-W. Lin, C. Zhang, A. Cano, “Exploring pattern mining algorithms for hashtag retrieval problem,” *IEEE Access*, vol. 8, pp. 10569–10583, 2020.
- [17] A. Karlekar, A. Seal, O. Krejcar, C. Gonzalo-Martin, “Fuzzy k-means using non-linear s-distance,” *IEEE Access*, vol. 7, pp. 55121–55131, 2019.
- [18] A. Seal, A. Karlekar, O. Krejcar, C. Gonzalo-Martin, “Fuzzy c-means clustering using jeffreys-divergence based similarity measure,” *Applied Soft Computing*, vol. 88, p. 106016, 2020.
- [19] K. K. Sharma, A. Seal, “Spectral embedded generalized mean based k-nearest neighbors clustering with s-distance,” *Expert Systems with Applications*, vol. 169, p. 114326, 2021.
- [20] K. K. Sharma, A. Seal, A. Yazidi, A. Selamat, O. Krejcar, “Clustering uncertain data objects using jeffreys-divergence and maximum bipartite matching based similarity measure,” *IEEE Access*, vol. 9, pp. 79505-79519, 2021.
- [21] A. Seal, E. Herrera Viedma, et al., “Performance and convergence analysis of modified c-means using jeffreys-divergence for clustering,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 2, pp. 141-149, 2021.
- [22] K. K. Sharma, A. Seal, A. Yazidi, O. Krejcar, “A new adaptive mixture distance-based improved density peaks clustering for gearbox fault diagnosis,” *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–16, 2022, doi: 10.1109/TIM.2022.3216366.
- [23] T. Ullmann, C. Hennig, A.-L. Boulesteix, “Validation of cluster analysis results on validation data: A systematic framework,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, p. e1444, 2022.
- [24] B. Tavakkol, J. Choi, M. K. Jeong, S. L. Albin, “Object-based cluster validation with densities,” *Pattern Recognition*, vol. 121, p. 108223, 2022.
- [25] K. K. Sharma, A. Seal, “Multi-view spectral clustering for uncertain objects,” *Information Sciences*, vol. 547, pp. 723-745, 2020.
- [26] K. K. Sharma, A. Seal, “Outlier-robust multi-view clustering for uncertain data,” *Knowledge-Based Systems*, vol. 211, p. 106567, 2021.
- [27] K. K. Sharma, A. Seal, E. Herrera-Viedma, O. Krejcar, “An enhanced spectral clustering algorithm with s-distance,” *Symmetry*, vol. 13, no. 4, p. 596, 2021.
- [28] B. Liang, J. Cai, H. Yang, “A new cell group clustering algorithm based on validation & correction mechanism,” *Expert Systems with Applications*, vol. 193, p. 116410, 2022.
- [29] H. Cui, M. Xie, Y. Cai, X. Huang, Y. Liu, “Cluster validity index for adaptive clustering algorithms,” *IET Communications*, vol. 8, no. 13, pp. 2256–2263, 2014.
- [30] B. Tang, S. Kay, H. He, “Toward optimal feature selection in naive bayes for text categorization,” *IEEE transactions on knowledge and data engineering*, vol. 28, no. 9, pp. 2508–2521, 2016.
- [31] S. Sra, “Positive definite matrices and the s-divergence,” *Proceedings of the American Mathematical Society*, vol. 144, no. 7, pp. 2787–2797, 2016.
- [32] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Fofou, A. Bouras, “A survey of clustering algorithms for big data: Taxonomy and empirical analysis,” *IEEE transactions on emerging topics in computing*, vol. 2, no. 3, pp. 267–279, 2014.
- [33] S. Sharma, “Applied multivariate techniques, jhonn wiley & sons inc.; 116, new york,” *Lewis-Beck vd*, vol. 1994, pp. 112–113, 1996.
- [34] L. Hubert, P. Arabie, “Comparing partitions,” *journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.
- [35] T. Caliński, J. Harabasz, “A dendrite method for cluster analysis,” *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

- [36] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.
- [37] D. L. Davies, D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.
- [38] A. B. Said, R. Hadjidj, S. Fougou, "Cluster validity index based on jeffrey divergence," *Pattern Analysis and Applications*, vol. 20, no. 1, pp. 21–31, 2017.
- [39] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.
- [40] U. Maulik, S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 12, pp. 1650–1654, 2002.
- [41] X. L. Xie, G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 13, no. 8, pp. 841–847, 1991.
- [42] M. Bouguessa, S. Wang, H. Sun, "An objective approach to cluster validation," *Pattern Recognition Letters*, vol. 27, no. 13, pp. 1419–1430, 2006.
- [43] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Perez, I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243–256, 2013.
- [44] S. Sra, R. Hosseini, "Conic geometric optimization on the manifold of positive definite matrices," *SIAM Journal on Optimization*, vol. 25, no. 1, pp. 713–739, 2015.
- [45] S. Chakraborty, S. Das, "k- means clustering with a new divergence-based distance metric: Convergence and performance analysis," *Pattern Recognition Letters*, vol. 100, pp. 67–73, 2017.
- [46] C. De Stefano, M. Maniaci, F. Fontanella, A. S. di Freca, "Reliable writer identification in medieval manuscripts through page layout features: The "avila" bible case," *Engineering Applications of Artificial Intelligence*, vol. 72, pp. 99–110, 2018.
- [47] D. Dheeru, E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: <http://archive.ics.uci.edu/ml>.
- [48] S. Affeldt, L. Labiod, M. Nadif, "Spectral clustering via ensemble deep autoencoder learning (sc-eda)," *Pattern Recognition*, vol. 108, p. 107522, 2020.
- [49] S. Wold, K. Esbensen, P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.
- [50] C. Ding, X. He, "K-means clustering via principal component analysis," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 29.
- [51] L. Bai, X. Cheng, J. Liang, H. Shen, Y. Guo, "Fast density clustering strategies based on the k-means algorithm," *Pattern Recognition*, vol. 71, pp. 375–386, 2017.



Krishna Kumar Sharma

He received PhD from the Computer Science and Engineering Department, PDPM Indian Institute of Information Technology, Design and Manufacturing Jabalpur, Madhya Pradesh, India in 2021 and has received the M.Tech.(Information Technology) degree from IIT Allahabad, Uttar Pradesh, India, in 2011. He is currently an Assistant Professor with the Computer Science and

Informatics Department, University of Kota, Kota, Rajasthan, India. His current research interest includes pattern recognition.



Ayan Seal

He received a Ph.D. in engineering from Jadavpur University, West Bengal, India, in 2014. He visited the Universidad Politecnica de Madrid, Spain as a visiting research scholar. He is currently an Assistant Professor with the Computer Science and Engineering Department, PDPM Indian Institute of Information Technology, Design and Manufacturing Jabalpur, Jabalpur, Madhya Pradesh,

482005, India. He is the recipient of several awards. He is at the top %2 scientists according to Stanford University, 2022. Dr. Seal has been granted

sponsored projects by the Govt. of India funding agencies. He has authored or co-authored several journals, conferences, and book chapters on the applications of computer vision. He is on the editorial board of several journals. His current research interests include computer vision, machine learning, deep learning, and brain-computer interface.



Anis Yazidi

He received the M.Sc. and Ph.D. degrees from the University of Agder, Grimstad, Norway, in 2008 and 2012, respectively. He was a Researcher with Teknova AS, Grimstad, Norway. From 2014 to 2019, he was an Associate Professor with the Department of Computer Science, Oslo Metropolitan University, Oslo, Norway, where he is currently a Full Professor, leading the research group in applied artificial intelligence. He is also Professor II with the Norwegian University of Science and Technology (NTNU), Trondheim, Norway. His current research interests include machine learning, learning automata, stochastic optimization, and autonomous computing.



Ondrej Krejcar

He is a full professor in systems engineering and informatics at the University of Hradec Kralove, Faculty of Informatics and Management, Center for Basic and Applied Research, Czech Republic; and Research Fellow at Malaysia-Japan International Institute of Technology, University Technology Malaysia, Kuala Lumpur, Malaysia. In 2008 he received his Ph.D. title in technical cybernetics at Technical University of Ostrava, Czech Republic. He is currently a vice-rector for science and creative activities of the University of Hradec Kralove from June 2020. At present, he is also a director of the Center for Basic and Applied Research at the University of Hradec Kralove. In years 2016-2020 he was vice-dean for science and research at Faculty of Informatics and Management, UHK. His h-index is 23 according Web of Science, with more than 2500 citations received in the Web of Science, where more than 150 IF journal articles is indexed in JCR index (h-index 27 at SCOPUS with more than 3200 citations). In 2018, he was the 14th top peer reviewer in Multidisciplinary in the World according to Publons and a Top Reviewer in the Global Peer Review Awards 2019 by Publons. Currently, he is on the editorial board of the MDPI Sensors IF journal (Q1/Q2 at JCR), and several other ESCI indexed journals. He is a Vice-leader and Management Committee member at WG4 at project COST CA17136, since 2018. He has also been a Management Committee member substitute at project COST CA16226 since 2017. Since 2019, he has been Chairman of the Program Committee of the KAPPA Program, Technological Agency of the Czech Republic as a regulator of the EEA/Norwegian Financial Mechanism in the Czech Republic (2019-2024). Since 2020, he has been Chairman of the Panel 1 (Computer, Physical and Chemical Sciences) of the ZETA Program, Technological Agency of the Czech Republic. Since 2014 until 2019, he has been Deputy Chairman of the Panel 7 (Processing Industry, Robotics, and Electrical Engineering) of the Epsilon Program, Technological Agency of the Czech Republic. At the University of Hradec Kralove, he is a guarantee of the doctoral study program in Applied Informatics, where he is focusing on lecturing on Smart Approaches to the Development of Information Systems and Applications in Ubiquitous Computing Environments. His research interests include Technical Cybernetics, Ubiquitous Computing, Control Systems, Smart Sensors, Wireless Technology, Biomedicine, Image Segmentation and Recognition, Biometrics, Biotelemetric System Architecture (portable device architecture, wireless biosensors), development of applications for mobile / remote devices with use of remote or embedded biomedical sensors.

Explaining Query Answers in Probabilistic Databases

Hichem Debbi*

Department of Computer Science, University of M'sila, M'sila (Algeria)

Received 8 August 2021 | Accepted 13 January 2023 | Published 24 July 2023



ABSTRACT

Probabilistic databases have emerged as an extension of relational databases that can handle uncertain data under possible worlds semantics. Although the problems of creating effective means of probabilistic data representation as well as probabilistic query evaluation have been addressed so widely, low attention has been given to query result explanation. While query answer explanation in relational databases tends to answer the question: why is this tuple in the query result? In probabilistic databases, we should ask an additional question: why does this tuple have such a probability? Due to the huge number of resulting worlds of probabilistic databases, query explanation in probabilistic databases is a challenging task. In this paper, we propose a causal explanation technique for conjunctive queries in probabilistic databases. Based on the notions of causality, responsibility and blame, we will be able to address explanation for tuple and attribute uncertainties in a complementary way. Through an experiment on the real-dataset of IMDB, we will see that this framework would be helpful for explaining complex queries results. Comparing to existing explanation methods, our method could be also considered as an aided-diagnosis method through computing the blame, which helps to understand the impact of uncertain attributes.

KEYWORDS

Causality, Conjunctive Queries, Explanation, Probabilistic Databases, Query Answers.

DOI: 10.9781/ijimai.2023.07.005

I. INTRODUCTION

RECENT applications in different domains are producing a huge volume of imprecise or uncertain data. Applications such as RFID and sensor networks are reporting frequently imprecise information, which is due mainly to measurement errors. In other applications such as information extraction from text or web pages, the extraction process yields automatically probabilistic results, due to the imprecise data source, or ambiguity in natural language text. A most known example is NELL [1], the Never Ending Language Learner that learns over time by reading the web [2]. It produces a set of facts, each one is assigned a probability score representing the confidence of the fact extracted. Business analysis, data cleaning, and integration are also quite active domains for dealing with uncertain data.

Due to the high importance of dealing with uncertain data, and since traditional databases do not have the ability to store and query uncertain data, probabilistic databases have emerged to address this issue. In probabilistic databases, the tuple may exist with a certain probability, or the values of some attributes may be uncertain [3]. We refer to both two types of uncertainty as *tuple-level uncertainty* and *attribute-level uncertainty*. For modeling both types, we use the *possible worlds semantics* [4]. It states that the real database is not known with certainty, therefore we introduce a probability distribution on all possible instances or worlds of this database.

Among major challenges in relational databases we find the following related problems: consistent query evaluation, data repairs, data cleaning, and explanation of unexpected query results. Consistent query evaluation (CQE) refers to computing meaningful answers to

queries when dealing with an inconsistent database [5]–[7]. A database is considered inconsistent if it does not satisfy a set of specifications called integrity constraints. To restore consistency with regard to these constraints, different database repairs semantics have been proposed. The general idea is based on finding a consistent database close to the inconsistent one with a minimal number of repairs, and look for answers that are true in all repairs [8]. When we want to deal with this inconsistency we can employ diagnostic approaches that try to find the root causes for this inconsistency. Thus we face a matter of causality. Similarly, in data diagnosis or in data cleaning [9], we look generally for a set of causes. However, we face more a question of causality when we try to explain unexpected query results. In this regard, many approaches have been proposed leading to an interesting subject, which is causality query answering. These approaches are based mainly on analyzing the query result in the form of the query lineage. Lineage is a standard and powerful tool that helps us to track every output tuple in the query result to its origins or input tuples.

Although uncertain data representation as well as querying have been addressed so widely [4], [10]–[12], low attention has been given to query answer explanation comparing to classical databases. Kanagal and Deshpande [13] addressed this issue into two dimensions: the qualitative dimension, which refers to a classical question, why such an output tuple is in the query result, thus they try to identify the cause of the answer; the quantitative dimension: why does an output tuple have a high probability, thus they try to identify the input tuples' probabilities that significantly contributed to the output probability. Measuring such contributions is based on measuring the probability difference when altering the probabilities of these input tuples.

In probabilistic databases, depending only on lineage does not provide enough explanation, since the context is quantitative and the query result usually consists of multiple tuples. In this regard, Re and Suciu [14] proposed two approximate lineage techniques, sufficient

* Corresponding author.

E-mail address: hichem.debbi@univ-msila.dz

and polynomial lineages that provide a high compact representation of lineage that only takes into account the influencing set of tuples. Providing an efficient small lineage through these approximate techniques helps to track down any derivations and provides better explanations. While this work considers only conjunctive queries, Kanagal et al. [13] addressed in addition aggregation and top-k queries. One additional difference is that the first one addresses the creation of lineage, the later build their algorithms on top of the lineage formula, and then try to extract the most influential input tuples. In contrast to looking for previous approaches that look for causes, Ceylan et al. [15] have investigated the explanation of probabilistic queries with the aim of finding the most probable database and the most probable hypothesis for a given query. They studied these problems with respect to both conjunctive and ontology-mediated queries, and the complexity analysis results showed that it could be helpful for applications in prediction and diagnosis tasks.

Causality plays an important role in explaining any phenomena. Halpern and Pearl have introduced a causality model, which they refer to as structural equations [16], [17]. This model of causality has been successfully used in many research areas. Based on this definition, Halpern and Chokler [18] introduced the definition of responsibility. Responsibility extends the concept of all-or-nothing of the actual cause $X = x$ for the truth value of a Boolean formula ϕ in (M, u) , where u refers to a context, and M refers to the causality model. It measures the number of changes that have to be made in u in order to make ϕ counterfactually depends on X . When we have an uncertainty around the context, we face in addition to the question of responsibility the question of blame.

In this paper, we employ these definitions: causality, responsibility and blame to provide explanations for query answers in probabilistic databases. We take the lineage produced as input, and then try to identify the most responsible tuples and the uncertain variables that have the most blame for such an outcome. Our work for identifying and ranking causes with their degree of responsibility is similar to previous works [19], [20] in a way that it tries to identify the most responsible tuples for such an outcome. However, we address them in an uncertain setting, which leads us to propose an extended version of responsibility, which is probabilistic responsibility. The second part of our work aims to identify the attributes having the most blame. Although the uncertainty in probabilistic databases is mainly associated with uncertain attributes, to our knowledge, there has not been an attempt before to address the contribution of uncertain attributes.

To clarify the role of such diagnostic information, we use in our paper the form of U-relational databases [21] that are based on attribute-uncertainty. A U-relational database is featured in many probabilistic database management systems such as MayBMS [22]. MayBMS is considered one of the most successful probabilistic systems, which is built on top of an existing relational database management system [23].

Our technique is complete in a way that it could touch both forms of probabilistic databases (tuple and attribute uncertainty) in a complementary way. While probabilistic responsibility is used for measuring the contribution of most responsible tuples, this same measure is employed then to rank uncertain attributes with the most blame. To our knowledge, this is the first attempt to employ all these notions together: causality, responsibility and blame in probabilistic databases. Our work is similar to Kanagal and Deshpande [13] in a way that the algorithm proposed is built on top of lineage. To show the effectiveness of our approach, we conducted two main experiments. We first evaluate the execution time of the proposed algorithm by varying a number of parameters. This experiment has been done on synthetic data. Then, we show the usefulness of our approach on a real probabilistic database, which is the IMDB database [24].

In the following we summarize the main contributions of the paper:

- We provide a causal-based explanation framework for analyzing the query answers in probabilistic databases.
- It is the first time, where all the notions of causality, responsibility and blame are combined together in a synergistic way.
- By performing extensive experiments, we will show that our framework is scalable even for large databases, inducing millions of causes.
- Although we do not have enough available real probabilistic datasets for evaluation, we succeeded to get promising results by executing our framework on a real-dataset, which is the IMDB dataset. To our knowledge, explaining queries over IMDB dataset has been addressed for the first time.
- Our method does not act just as an explanation method for query answers, but also acts as an aided diagnostic method that helps to understand the contribution of uncertain attributes.
- The experiment on IMDB has been executed on the top of one of the successful probabilistic database management systems MayBMS [23].

The rest of this paper is organized as follows. In Section 2 we present some related works. We present some preliminaries and definitions in Section 3. We introduce U-relational databases as well as lineage in this section. In addition, we revise the definitions of causality and responsibility in relational databases. In Section 4, we present our definitions for causality, responsibility and blame in the context of probabilistic databases. This section is ended by introducing an algorithm for computing all the measures related to our definitions. Experimental results are presented in Section 5. The last section concludes the paper and outlines some future work.

II. RELATED WORK

Meliou et al. [19], [25] have addressed causality in relational database for the first time based on the definition of causality by Halpern and Pearl [16]. Given a query output over a database instance, they try to find the responsible tuples that cause this answer. A tuple t is considered as a cause for a query answer, if there exists such a contingency that represents a set of tuples called endogenous, in way that removing this set from the database makes the query output counterfactually depends on t , i.e, removing t will lead to a non-answer. This definition has been related to lineage based on *c-tables* [26], [27]. The lineage formula is introduced in Disjunctive Normal Form (DNF), in which, every tuple is represented by a Boolean variable, then a cause is considered as a tuple associated with a Boolean variable that is included in a minterm of the lineage formula. This definition has been enhanced by another quantitative measure called responsibility, which measures the degree to which a tuple is considered as a cause [18]. Computing responsibility of a tuple t is based on the size of its contingency, where the tuple with the lower contingency is supposed to has the highest responsibility and vice versa. In an extended work [25], they introduced a careful complexity analysis of computing causality and responsibility in databases for both why so and why no (non-answer) causality. A non-answer query result refers to the question: why some tuples are missed from the output? In this regard, Diestelkämper et al. [28] addressed this issue in the context of Big data, on queries of the data-intensive scalable computing (DISC) Appach spark. Despite the application context of Appach Spark, which is so novel and relevant, this work compared to previous works on missing answers, addresses nested data, which lead to rely on specific a nested data model and nested relational algebra for bags as a query language. However, this work just like previous works on missing answers analysis relies on lineage or provenance, specifically the why-not provenance.

Some reported open problems by [19] [25] have been addressed later by Salimi [29]. Among problems addressed are databases repairs and consistent query answering, abductive reasoning in databases, and the view-update problem. They argue that these famous problems in databases have strong connections to the definition of causality in databases. They showed that computing causes and their responsibilities can be considered as a consistent query answering problem, and in order to obtain repairs, they proposed algorithms for computing causes and their responsibilities for queries as unions of conjunctive queries [30]. Based on Hitting sets and vertex covers problems in hyper-graphs, they introduced more details on the complexity analysis of causality in databases, that is uncovering some complexity issues that have not been addressed by Meliou et al. [25]. They also showed the connection between query answer causality and abductive diagnosis as well as view-update problem, particularly, the delete-propagation problem where only tuple deletions are allowed from views [20]. Delete-propagation process tries to minimize the side-effect on views, thus, looking for a minimal set of tuples deletion. This issue has been related to a minimal contingency set of a causes to establish the connection between the two concepts. In addition, for establishing this connection, they adapted conditioning causality [31] that states that computing causes for an unexpected answer should be guided or conditioned on some prior knowledge of the correctness of another set of outputs. As a result, a measure for a tuple contribution to different outputs may be obtained. They also showed the connection between abductive diagnosis and query answering causality in the context of Datalog queries.

Another interesting work that employs causality and responsibility has been done by Lian and Chen [32]. They addressed the problem of *probabilistic nearest neighbor* (PNN), which is related the context of moving objects such as RFID, sensor networks and location-based services that introduce usually imperfect estimations of objects positions. In this work, responsibility is assigned to an object. This object is considered as a cause for other objects to be or not to be included in PNN query answers.

In addition to relational databases, causality and responsibility have been implemented in knowledge based systems. Mu [33] has addressed the inconsistency of knowledge bases through affecting a degree of responsibility for each formula, starting from the hypothesis that this responsibility should be explained from a causal perspective. In this setting, computing the degree of responsibility of a formula for inconsistency is based on identifying the minimal number of formulas that have to be removed from the knowledge base in order to break all the minimal inconsistent subsets not containing the formula, where the formula is declared not responsible if it does not belong to any minimal inconsistent subset.

In contrast to all previous works adopting lineage for both traditional and probabilistic databases, Miao et al. [34] proposed CAPE (Counterbalancing with Aggregate Patterns for Explanations) for explaining aggregation queries that does not rely on lineage or provenance at all. They consider that in the presence of outliers in data, relying on lineage only could be misleading. Therefore, they look for some patterns that hold on the data to provide explanations counterbalancing the user's observation. That is, outliers contradicting some pattern related to the user question might be easily identified.

III. PRELIMINARIES

A. Probabilistic Databases

In probabilistic databases, a database instance could be in several states, where each state has a degree of uncertainty. That is, we could have several possible instances, called worlds, each of which has a

probability. To model these instances under uncertainty we use the possible worlds semantics [4]. It states that a probabilistic database is represented as a finite set of possible worlds with some weights, and these weights sum up to 1. We refer to such a finite set of structures an *incomplete database* [2].

Let us fix a relational schema that consists of k relation names R_1, R_2, \dots, R_k . We refer to an incomplete database as $W = \{W^1, W^2, \dots, W^n\}$, where each $W^i = \langle R_1^i, R_2^i, \dots, R_k^i \rangle$ is a database instance. Now we define a probabilistic database as follows:

Definition 1. Probabilistic Database. A Probabilistic database is a probabilistic space $D = (W, P)$ over an incomplete database W , where $P: W \rightarrow [0,1]$ is a probability distribution function such that $\sum_{w \in W} P(W) = 1$.

In probabilistic databases, relations instances are supposed to be different from a world to another. We call a relation R_j certain or deterministic, if $R_j^1 = R_j^2 = \dots = R_j^n$. Relations are different in the way that each relation instance R_j^i in a world W^i contains different tuples from an instance of the same relation in another world. These tuples are considered as probabilistic events because we are not sure that they represent certain data. Therefore, we define for each tuple a probability called *marginal probability* or *confidence*. The probability of a tuple $t \in R_j$ is defined as follows:

$$P(t \in R_j) = \sum_{1 \leq i \leq n: t \in R_j^i} P(W^i) \quad (1)$$

The question that arises now is how to evaluate a query Q on a probabilistic database. For doing so, two semantics have been considered so far. The first is *possible answer sets* semantics, where the query is evaluated on every possible world, this returns a set of tuples for each world. Since the representation of all answers is not practical, we instead use the second semantics, which is called *possible answers* [2]. As in the first semantics, the query is evaluated on all possible worlds, however, the result is returned as a list of tuples annotated with probabilities. We can say that such a tuple t is a possible answer to a query Q , if $\exists W \in \mathbf{W}$ such that $t \in Q(W)$. We can say that a tuple is certain if $\forall W \in \mathbf{W}$, $t \in Q(W)$. Given a query Q and a probabilistic database $D = (W, P)$, the marginal probability of a tuple $t \in Q$ is $P(t \in Q) = \sum_{W \in \mathbf{W}: t \in Q(W)} P(W)$. That is, the marginal probability of a tuple t is computed by summing up the probabilities of worlds in which the tuple t is returned as an answer for the query Q . So, the possible answers for a query Q with their probabilities are represented as $Q(D) = \{(t_1, p_1), (t_2, p_2), \dots\}$. We can easily notice that two variants of tuple answer semantics can be defined, possible and certain. These sets are defined as follows :

$$Q_{\text{poss}}(\mathbf{W}) = \{t | (t, p) \in Q(D), p > 0\} \quad (2)$$

$$Q_{\text{cert}}(\mathbf{W}) = \{t | (t, p) \in Q(D), p = 1\} \quad (3)$$

B. BID Database

Probabilistic databases can be obtained through two types of uncertainties, either on the level of tuples, this type is called *tuple-level uncertainty*, or on the level of attributes, and this is called *attribute-level uncertainty*. In the first type, a tuple is considered as a random variable, so given a database instance, we are not sure if this tuple really exists. In the second type, the values of specific attributes are uncertain for each tuple, that is, the attribute is considered as a random variable, its domain is all the values that the attribute may take. During the query evaluation process, attribute-level uncertainty is usually transformed to a tuple-level uncertainty. Based on these types of uncertainty, we have two types of probabilistic databases: tuple-independent database, where the tuples are independent probabilistic events, and block independent-disjoint probabilistic database, where the tuples are partitioned into blocks according to an uncertain attribute, such that

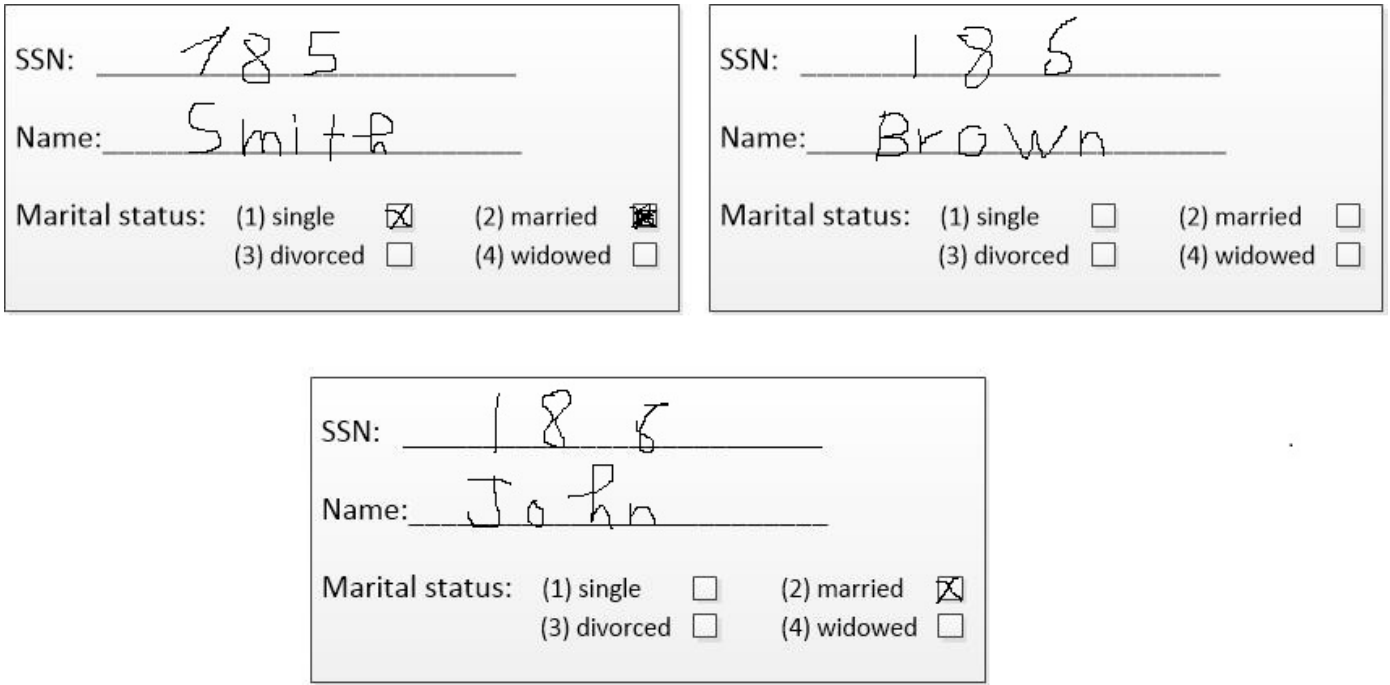


Fig. 1. Survey Forms.

tuples in the same block are disjoint and their probabilities sum up to 1, and tuples from different blocks are independent. So, in BID database, it is not possible for a world to contain more than one tuple from the same block. This representation is very effective and considered as a complete representation system with conjunctive query views [2]. This representation has been featured in many probabilistic database systems, such as MayBMS [22], [35], MystiQ [36] and Trio [37],[38]. It can also help for capturing key violations in databases.

Consider the forms presented in Fig. 1. It corresponds to 3 survey forms filled by three persons: Smith, Brown and John. Each form contains as information: social security number (SSN), name, and marital status (M). While both attributes ID and name have evident values, it is not the same case for SSN and M. For instance, one could be mistaken for the values of SSN in form 1, whether it is 185 or 785, and also in form 2 (185/186). One could be also mistaken for the value of M in form 1, whether its value is "single" or "married", because it seems that Smith first checked "single" mistakenly, then he checked "married", but it could also be the contrary. In the second form, John did not check any status, hence, we have have four possible values. An example of BID database regarding these forms is presented in Fig 2. By taking exactly one value of each uncertain variable, this could result in $2 \times 2 \times 2 \times 3 = 24$ possible readings of the three forms. Suppose that we have millions of forms filled with this uncertainty, it would be a very challenging task to represent and process all these possible readings.

Fid	SSN	P	Fid	M	P
1	185	0.4	1	1	0.7
1	785	0.6	1	2	0.3
2	185	0.7	2	1	0.25
2	186	0.3	2	2	0.25
3	186	0.75	2	3	0.25
3	188	0.25	2	4	0.25
			3	1	1

Fig. 2. An example of BID database.

C. U-Relational Database

U-relational database [11] is based on BID representation, and it is also considered as a probabilistic extension of classical conditional tables (C-tables) [39]. In a c-table, each tuple is annotated with a propositional formula over random variables. Using the logical operations *And* (\wedge), *OR* (\vee) and *NOT* (\neg), the propositional formula is obtained. In a U-relational database, the schema of each U-relation consists of: a tuple id column, a set of column pairs (V_p, D_p) that represent variable assignments or valuations, and finally a set of value columns. The probabilities of the assignments are stored in a separate table $W(V, D, P)$, called the world table.

Definition 2. U-relation Schema. A U-relation schema is a represented as $S = (V_1, D_1, \dots, V_k, D_k, A_1, \dots, A_m)$, where k refers to the number of pairs of variable assignments (distinguished attributes), and m refers to the number of value columns. A U-relational database consists of U-relations.

A U-relational database is an efficient and complete representation system that could provide a compact representation of the exponential number of possible worlds, and could also allow the representation of the result of any query [21]. Given U-relation database, the result of a query can be also returned in the same representation. A U-relational database for the example presented in Fig. 1 is presented in Fig. 3

In U-relational databases, each world W is defined by an assignment θ that assigns one possible value to each variable. Then, the probability or weight of this possible world is computed as the product of the probabilities associated to these valuations. For instance the probability of the world $W = \{x \mapsto 1, y \mapsto 3, z \mapsto 4, u \mapsto 1, v \mapsto 3, w \mapsto 1\}$ is $0.4 \times 0.3 \times 0.25 \times 0.7 \times 0.25 \times 1 = 0.0056$.

D. Lineage and Conjunctive Queries

Lineage is defined as propositional formula over input tuples in a database in order to explain how such an output query has been derived. Hence, each output tuple has a Boolean formula that states the input tuples responsible for its occurrence. Lineage is considered as a powerful tool for explaining query results. However, projection of million tuples on a single output tuple could result in a huge lineage formula. By considering uncertain context, i.e. probabilistic databases, the interpretation of lineage is more challenging. Lineage

$U_R[SSN]$	$V \mapsto D$	Fid	SSN
	$x \mapsto 1$	1	185
	$x \mapsto 2$	1	785
	$y \mapsto 1$	2	185
	$y \mapsto 3$	2	186
	$z \mapsto 3$	3	186
	$z \mapsto 4$	3	188
$U_R[M]$	$V \mapsto D$	Fid	M
	$u \mapsto 1$	1	1
	$u \mapsto 2$	1	2
	$v \mapsto 1$	2	1
	$v \mapsto 2$	2	2
	$v \mapsto 3$	2	3
	$v \mapsto 4$	2	4
	$w \mapsto 2$	3	2

$U_R[Name]$	Fid	Name	W	$V \mapsto D$	P
	1	Smith		$x \mapsto 1$	0.4
	2	Brown		$x \mapsto 2$	0.6
	3	John		$y \mapsto 1$	0.7
				$y \mapsto 3$	0.3
				$z \mapsto 3$	0.75
				$z \mapsto 4$	0.25
				$u \mapsto 1$	0.7
				$u \mapsto 2$	0.3
				$v \mapsto 1$	0.25
				$v \mapsto 2$	0.25
				$v \mapsto 3$	0.25
				$v \mapsto 4$	0.25
				$w \mapsto 2$	1

Fig. 3. An example of U-relational database.

in probabilistic databases is defined in a similar way comparing to relational databases, furthermore, query evaluation reduces to computing the lineage of output tuples, and then computing the probabilities of the lineage formulas.

Actually, all probabilistic database management systems use lineage-based query evaluation [12]. Many techniques have been proposed for computing the probability of propositional formulas in an efficient way, such as read-once formulas [40], OBDD [41], and d-DNNF [42].

Lineage formulas are written usually as DNF formulas. DNF formulas are Boolean formulas in disjunctive normal form. Formally, given a query Q and a probabilistic database D , each answer tuple $a \in Q(D)$ is associated with a DNF formula $\lambda_a^{Q,D}$.

In U-relational database systems like MayBMS, computing the probability or confidence of an output tuple is based on computing the probability of a DNF of this tuple, by summing up the probabilities identified with the valuation θ of random variables such that DNF becomes true under θ . Some approximation techniques based on Monte Carlo simulation have been proposed to approximate DNF probabilities computation [4], [43]. One interesting point about U-relational databases, is that a lineage of output tuples can be written in k-DNF formula for small number of k , where k represents the maximum number of literals.

Conjunctive query. Conjunctive query is the simplest and most used query in relational databases. It is restricted to the operators \exists and \wedge and it has the following form in relational calculus: $x, y \mid \exists z (R(x, y) \wedge S(y, z))$ where R and S are two relations. In Datalog, it is given in the form: $q(x, y) : -R(x, y), S(y, z)$. In SQL, conjunctive queries correspond to *select - project - join*, and the *where* clause

contains only equalities. By relating conjunctive queries to lineage, in SQL we can use: *select distinct attributes*, or alternatively the following form: *select * order by attributes*.

The lineage of an output tuple of a query in the first form would be a DNF formula whose terms are given by the rows returned by a query in the second form.

Example III.1. Consider the U-relational database in Fig. 3. Consider a conjunctive query on two uncertain relations $U_R[SSN]$ and $U_R[M]$ that returns all possible SSNs of married persons (*select SSN from U_SSN, U_M where U_SSN.FID = U_M.FID and U_M.M=2*). Before going through query results, we should mention that some resulting worlds of this database could be inconsistent, in way that one world could contain two persons with the same SSN, which is not possible. Database management systems like MayBMS offers the possibility to detect such violations. An example of a query that repairs this database is given as the following:

repair key (fid) in Census_SSN weight by p;

Where *Census_SSN* refers to the relation *URrSSNs*. After applying this constraint, we can reduce the number of instances of *URrSSNs* to four possible instances. Although the number of possible worlds is obviously more than 4, because we have another uncertain relation in hand, which is $U_R[M]$, we fix four instances as well for this relation to better explain our future notions. That is, we consider only the following 4 worlds:

$$\begin{aligned}
 W_1 &= \{x \mapsto 1, y \mapsto 3, z \mapsto 4, u \mapsto 2, v \mapsto 2, w \mapsto 2\} \\
 P(W_1) &= 0.4 \times 0.3 \times 0.25 \times 0.3 \times 0.25 \times 1 = 0.0022 \\
 W_2 &= \{x \mapsto 2, y \mapsto 1, z \mapsto 4, u \mapsto 1, v \mapsto 2, w \mapsto 2\} \\
 P(W_2) &= 0.6 \times 0.7 \times 0.3 \times 0.7 \times 0.25 \times 1 = 0.022 \\
 W_3 &= \{x \mapsto 2, y \mapsto 3, z \mapsto 4, u \mapsto 2, v \mapsto 1, w \mapsto 2\} \\
 P(W_3) &= 0.6 \times 0.3 \times 0.25 \times 0.3 \times 0.25 \times 1 = 0.0033 \\
 W_4 &= \{x \mapsto 2, y \mapsto 1, z \mapsto 3, u \mapsto 1, v \mapsto 1, w \mapsto 2\} \\
 P(W_4) &= 0.6 \times 0.7 \times 0.75 \times 0.7 \times 0.25 \times 1 = 0.055
 \end{aligned}$$

For a query that returns married persons ($M = 2$), the query returns over these 4 worlds 8 possible tuples. Thus, the lineage formula would be a DNF formula consisting of 8 terms returned by this query over 4 possible worlds

$$\begin{aligned}
 &[(x = 1 \wedge u = 2) \vee (y = 3 \wedge v = 2) \vee (z = 4 \wedge w = 2)] \vee \\
 &[(y = 1 \wedge v = 2) \vee (z = 4 \wedge w = 2)] \vee [(x = 2 \wedge u = 2) \vee \\
 &(z = 4 \wedge w = 2)] \vee [(z = 3 \wedge w = 2)]
 \end{aligned}$$

The output probability is computed with respect to this DNF formula, which is the sum of these probabilities: 0.0825.

E. Causality and Responsibility in Relational Databases

Counterfactual reasoning that states: event A is a cause of event B if, had A not happened then B would not have happened, plays an important role in causality. Halpern and Pearl [16] extended this basic statement by taking A to be a cause of B, if B counterfactually depends on A under some contingency. Following this definition, Meliou et al. have introduced the definition of database causality. Let us fix a relational schema that consists of k relation names R_1, R_2, \dots, R_k . Given a database instance D and a conjunctive query Q , for each relation R_i we denote by R_i^n the set of endogenous tuples, and R_i^x the set of exogenous tuples. Endogenous tuples are those that are considered to be causes, whereas exogenous tuples are deemed not to be possible causes. That is, the tuples in D are partitioned into two sets $D = D^n \cup D^x$, where D^n, D^x represent all endogenous and exogenous tuples respectively. A tuple $t \in D^n$ is said to be a counterfactual cause for an answer a to Q in D , if $D \models Q(a)$ and $D - \{t\} \not\models Q(a)$. Given this definition, we can now give the definition of actual cause.

Definition 3. Actual cause. A tuple t is an actual cause for an answer a in D if there exists a set of endogenous tuples $\Gamma \subseteq D^n$, such that t is a

counterfactual cause for a in $D - \Gamma$. We call this set a contingency for t .

That is, it is sufficient to find a set of tuples Γ that can be removed in order to make the query answer counterfactually depends on the existence of t . In other words, removing t will switch to non-answer. It is evident that every counterfactual cause is an actual cause by taking $\Gamma = \emptyset$. Based on c-tables representation, computing the causes has been related to lineage in DNF formula [19]. Let us assume that every tuple t in D is associated with a Boolean variable X_t , and we denote by X^n the Boolean variables related to endogenous tuples. The DNF formula $\lambda_a^{Q,D}$ for such an output tuple a will consist only of X^n . In terms of lineage, we say that a tuple t is an actual cause for an answer a to a query Q on D , iff there exists a minterm in $\lambda_a^{Q,D}$ that contains t .

An open question has arisen given the above definition concerning a contingency Γ : does it matter the size of Γ on the importance of an actual cause? Chockler and Halpern have addressed this issue by introducing a quantitative measure called responsibility [18]. In relational databases responsibility has been used to rank tuples [19], [20].

Definition 4. Responsibility. *The degree of responsibility of a cause t for an answer a of a query Q on D , denoted dr_t , is $dr_t = 1/|\Gamma| + 1$, where Γ represents the minimal contingency set for t .*

That is, the degree of responsibility of a tuple t is based on computing the number of tuples that we need to remove from the database D in order to make a query answer a contractually depends on t . Obviously, when the actual cause t is already a counterfactual cause, i.e., $\Gamma = \emptyset$, then the degree of responsibility $dr_t = 1$.

Let us consider a simple conjunctive query $q_1(z)$: $-R(x, y), S(y, z)$ over a database that contains the tuples $R(a, b), S(b, d), R(a, c), S(c, d)$. The lineage for the answer d would be $(X_{R(a,b)} \wedge X_{S(b,d)}) \vee (X_{R(a,c)} \wedge X_{S(c,d)})$. It is evident that every tuple here is an actual cause for this answer. Removing $R(a, b)$ or $R(a, c)$ for instance will make both $S(b, d)$ and $S(c, d)$ counterfactual causes. It is evident that the minimal number of tuples needed for making these causes counterfactual is 1, that is each tuple here has a degree of responsibility $1/2$. Now let assume the $q_2(y)$: $-R(x, y), S(y)$ over a database that contains the tuples $R(a, c), R(b, c), S(c)$, the lineage for the answer c would be $(X_{R(a,c)} \wedge X_{S(c)}) \vee (X_{R(b,c)} \wedge X_{S(c)})$. Here the tuple $S(c)$ has a 1 as a degree of responsibility since its removal will result in no answer, whereas both of the tuples $R(a, c)$ and $R(b, c)$ share the responsibility.

IV. CAUSALITY AND RESPONSIBILITY IN PROBABILISTIC DATABASES

It is evident that causality and its quantitative extension responsibility play an important role in explaining query results in relational databases. Since probabilistic databases extend relational databases with probabilistic semantics, a great effort has been done mainly to extend relational semantics to represent uncertainty in data [2]. In this regard, we choose to extend the definitions of causality and responsibility to probabilistic databases, in order to explain the probabilistic results of a query over a probabilistic database. In addition, we use the definition of blame [18] that we consider as a helpful tool when we want to go deeply for explaining such results. Kanagal and Deshpande [13] have introduced some fundamental notions for explaining probabilistic results over probabilistic databases. They addressed questions like: why such a tuple is included in the result? in addition to: why a tuple t has more probability than the probability of another tuple t' ? Some works try to enhance the computed lineage itself by proposing the notion of *approximate lineage*, which is considered to include only the most important and influential tuples. Here, we rely on complete lineage, and since it could be very huge, we employ these two definitions (causality and responsibility) to guide

the user to the most responsible causes for such an output.

We should recall that probabilistic databases can be given into two representations: tuple-based and attribute-based. The approach that we are going to propose can handle both sources of uncertainties. While causality and responsibility can be used for explaining the first type, we count on blame for understanding the contribution of uncertain attributes to the probabilistic result of the query. So, we depend on causality and responsibility for answering questions like: Why is this tuple? what is its responsibility? what contribution does it have on the probability of the query? On the other side, we depend on blame for answering the question: *What uncertain variable we should blame the most for such an outcome?* While the first question has been considerably addressed, to our knowledge, the second question has not been addressed before, though it is of high importance and could return deeper explanations. Furthermore, it could be very helpful for understanding the entire design of probabilistic databases.

Here we define causality, responsibility and blame by considering the probabilistic database to be U-relational database. As we showed before, U-relational databases are a complete representation, effective in many ways, and more importantly, it explicitly features attribute-uncertainty, which will be a ground for our definitions. One more important feature of U-relational databases is that causes can be returned in a detailed manner, i.e., in a form of valuations of the uncertain variables, this could provide a better explanation than an entire tuple that could consist of many attributes.

We introduce first the notions of causality and responsibility in general semantics, the possible worlds semantics. Hence, our definitions would be applicable for any probabilistic database, then we restrict our presentation with an example to U-relational database.

We should recall that computing the probability of a query output in U-relational databases is based on computing the probability of a DNF, which is the sum of the weights of the worlds identified with valuations θ such that the DNF becomes true under θ [44].

A. Causality and Responsibility

Before introducing our definitions, let us recall first what is a causal model according to Halpern and Pearl [16], on which the definitions of causality, responsibility and blame are built. A causal model is a tuple $M = (U, V, F)$, where the set U represents exogenous variables, whose values are determined by factors outside the model M , but they are necessary to encode the context, and the set of endogenous variables V , whose values are determined by a set of functions F . The causes are determined then by the valuations of V . The causality model can be extended to a probabilistic causal model as a tuple (M, Pr) , where M is the causality model, and Pr is a probability function over possible contexts [17].

To relate this definition to probabilistic databases, the context will refer to a possible world W , whose probability is defined by the probability function P . V will refer to endogenous tuples in this world, this set is included in the DNF formula. U will refer to exogenous tuples that are not included in the DNF formula, but they are necessary to define the possible world. By considering U-relational databases, F will refer to the valuations θ .

Now we can introduce the definition of a cause in probabilistic databases. Let us consider a U-relational database D . We should recall that a U-relation schema consists of variables assignments (V, D) , such that every tuple is associated with a valuation $V_i = v_i$. We recall also that given a query Q and a database D , each answer tuple $a \in Q(D)$ is associated with a global DNF formula $\lambda_a^{Q,D}$. Let us denote by $\lambda_a^{Q,W}$ the lineage for a in the local level of a world W .

Definition 5. Actual cause. *A tuple t with a valuation $V_i = v_i \in \lambda_a^{Q,W}$ is an actual cause for an answer a in a possible world W , subsequently*

in D , if there exists a subset of variables $\Gamma \subseteq V$, such that switching their values makes the truth value of $\lambda_a^{Q,W}$ counterfactually depends on $V_i = v_i$.

That is, this definition is the same for a cause in relational database, just by considering one possible world. V refers to the variables associated with endogenous tuples, and $V_i = v_i$ is a valuation associated with an input tuple.

Definition 6. Responsibility. The degree of responsibility of a tuple t with a valuation $V_i = v_i$ for an answer a in a possible world W , is $dr_t(V_i = v_i)^W = 1/|\Gamma|+1$, where Γ represents the minimal contingency set for t .

Conjunctions in lineage represent join over relations, that is, a valuation could appear in different conjunctions.

Proposition IV.1. Let $C_W = C_1, \dots, C_n$ be the set of conjunctions related to DNF formula $\lambda_a^{Q,W}$. A tuple t with a valuation $V_i = v_i$ is a counterfactual cause, and thus it has responsibility 1, iff t is included in every conjunction C_i .

That is the cause with the highest responsibility will be the one appearing the most in the conjunctions of $\lambda_a^{Q,W}$.

Although, a cause responsibility is a very interesting measure, in this simplest form it did not yet take into account the uncertainty introduced by random variables V . Therefore, we have to deal with this uncertainty. Halpern and Pearl [17] addressed also the case where the context is uncertain, then the probability of a cause $X = x$, where $X \in V$, is given by the probability of the context u on which X has its value. We adopt the same definition here for cause in probabilistic database, and we define the cause probability in each world W as:

$$Pr_t(V_i = v_i)^W \stackrel{def}{=} P(W) | t \in W \quad (4)$$

That is, we have associated for each cause a probability that represents exactly the probability of the world in which it is included. Let us now denote by $W_a^{Q,D} \in W$ the set of worlds related to the global DNF formula $\lambda_a^{Q,D}$. Let a tuple t with a valuation $V_i = v_i$ be a cause that appears in a set of worlds $W_t \subseteq W_a^{Q,D}$.

Given the definitions of cause responsibility and cause probability, we introduce the following definition.

Definition 7. Probabilistic responsibility. We define for each cause for an answer $a \in Q(D)$ a probabilistic responsibility as follows:

$$drP_t(V_i = v_i) = \sum_{W \in W_t} Pr_t(V_i = v_i)^W \times dr_t(V_i = v_i)^W \quad (5)$$

That is, we define for each cause a responsibility over possible worlds, where each cause takes the probability of the world in which it is included. This has been done by computing the product of the cause responsibility and cause probability on the local level for every world, then summing up these measures. We consider this measure as an enhanced and required version of classical responsibility.

Definition 8. Most responsible cause. A tuple t with a valuation $V_i = v_i$ is a most responsible cause for an answer $a \in Q(D)$, if $drP_t(V_i = v_i) \geq drP_{t'}(V_i = v_i)$.

Consider for instance a world where we have a cause with responsibility 1. According to the classical definition, this cause represents a good explanation, however, when we know that the probability of this world is very low, the cause automatically loses its importance. Therefore, most responsible causes are those having high responsibilities in worlds with high probabilities. Obviously, we are interested more in most probable worlds. We should note here that a most probable database is supposed to be the closest to a certain database. It is evident that a most responsible cause does not have necessarily the highest probability, and it might not be unique.

Proposition IV.2. A cause t could have $drP_t(V_i = v_i) = 1$ for an answer $a \in Q(D)$, which is the highest value possible for probabilistic causality,

iff $t \in Q_{cert}(W)$, and t is a counterfactual cause.

Proof. A certain tuple $t \in Q_{cert}(W)$ is defined as a tuple that is present in every world $W \in W$ -See equation (3)-, i.e, $P(t) = 1$, which represents the sum of all worlds probabilities. With respect to Definition 7, the probabilistic responsibility of a tuple t (drP_t) is obtained by the product of its probability (P_t) and responsibility dr_t , so a cause probability is weighted by its responsibility. So, if this tuple t is a counterfactual cause in every world, then it has always 1 as a degree of responsibility, and since $t \in Q_{cert}(W)$, then the probabilistic responsibility of t will represent exactly the sum of all worlds probabilities, which is certainly 1, i.e, $drP_t(V_i = v_i) = 1$.

B. Blame

Blame has been proposed as a complementary notion for responsibility in the presence of uncertainty [18]. Actually, blame is used when responsibility does not give enough explanation, one source for this incompleteness is the uncertainty associated with the contexts in which responsibility is measured. As presented before, causality and responsibility have been proved to be a helpful tool for explaining such an answer, but we still need to question the probabilities derived. How uncertain attributes have contributed to this result, more deeply, which uncertain variable has the most blame for such an outcome. This question has more importance with conjunctive queries involving many uncertain relations.

Definition 9. Blame. Let X be an uncertain attribute, and $U_R[X]$ the uncertain relation. Let us denote by $X_{r,W}$ the set of causes with valuations $V_i = v_i, \dots, v_m$ with respect to X in a world W . We define for each uncertain attribute X for an answer $a \in Q(D)$ the degree of blame as follows:

$$db(X) = \sum_{W \in W_t} \prod_{t \in X_{r,W}} P(V_i = v_i) dr_t(V_i = v_i)^W \quad (6)$$

So, the degree of blame for an uncertain attribute X is based on computing first the products of the tuples probabilities associated with it in the world table, and their responsibilities with respect to each world, then summing up all these measures. That is we are trying to quantify the probability contribution of each cause valuation to the probability of the world taking into account its responsibility. We can say that we obtain here two diagnostic information, the first concerns tuples (causality and responsibility), and the second concerns the uncertain attributes (blame), where the second notion is based on the former notions.

Theorem IV.3. If a tuple t with a valuation $V_i = v_i \in X_{r,W}$ is a most responsible cause, it does not mean necessarily that X has the most blame.

Proof. We can think of responsibility as an adjusting factor for both probability of a tuple P for computing the blame, and probability of cause Pr for computing the probabilistic responsibility. Let us consider that a cause t with a valuation $V_i = v_i$ is included in the set of certain tuples $Q_{cert}(W)$, that is according to proposition IV.2, this is a most responsible cause with the highest possible value. It is sufficient to find an uncertain attribute $X2$ such that $db(X2) > db(X1)$. We should notice that the most responsible cause when it comes to blame, is affected by the rest of valuations in $X_{r,W}$. That is, it can be found that $X1_{r,W}$ has low probabilities as well as low responsibilities compared to the valuations $X2_{r,W}$, where $X2_{r,W}$ does not include the most responsible cause.

C. An Algorithm for Explaining Query Answers

Our algorithm takes the lineage of such an answer in DNF form, and then returns the causes ordered with their probabilistic responsibilities, and the attributes ordered with their blame. However, before analyzing the DNF formula, some pre-processing is performed in order to compute the results in an efficient way. We propose to transform the DNF formula into a table of two dimensions, the first dimension represents the local lineage formula with respect

to each world, and the second represents the possible worlds. This representation provides all the information that we need to compute responsibility and blame. We call this representation a causality matrix.

The algorithm 1 aims to generate causes with their dR and Pr , and attributes with dB . To do so, we should introduce three main functions representing the main steps in computing probabilistic responsibility and blame. The objective of the first function *ComputeCauses* is to extract the causes from each term in the DNF formula, with respect to each world, the result will seem like a table of two dimensions. This table will contain redundant causes that appear in multiple terms, however, we need it to start the next phase. The second phase takes the result of the first phase as input, and tries to compute for each distinct cause its responsibility. This requires the computation of the contingency Γ for each cause. Based on the input, this is can be easily achieved through the use of a variable K . When the cause is not contained in a minterm, K is increased by one, which means that this term could be removed without affecting the truth value of $\lambda_a^{Q,D}$. Similarly, if the cause is contained in every term, thus, $K=0$, it means that this cause is counterfactual and has responsibility 1. By completing the two loops, each cause will have its related responsibility. After that, we can compute the probabilistic responsibility of each cause. By completing this phase, we will have all required information in hand (causes with their dR , Pr and P) to compute the blame.

Algorithm 1 Compute Explanations

- 1: **Inputs:** DNF formula $\lambda_a^{Q,D}$
 - 2: **Outputs:** Causes with responsibilities, attributes with blame
 - 3: $Causes = \emptyset$
 - 4: $Causes = \text{ComputeCauses}(\lambda_a^{Q,D})$
 - 5: **ComputeRespons(Causes)**
 - 6: **ComputeBlame(Causes with dr and Pr)**
 - 7: **OutputExplanations(Causes with dR and Pr, Attributes with dB)**
-

It is evident that this algorithm returns the set of explanations in polynomial time. Its complexity depends on the size of D and the size of the DNF formula. For computing responsibility, since the DNF formula is related to a linear conjunctive query, computing responsibility has been already proved that it can be achieved in polynomial time [25], [45]. They have proposed a careful analysis of complexity for both causality and responsibility in relational databases, for both *Why so?* and *Why no?* causality.

Example IV.1. Let us consider the results of the previous example of the census database (Example III.1). We want now to run our algorithm and measure the contribution of each cause to the previous output. We compute tuples responsibilities, as well as blame for the uncertain attributes SSN and M . We need first to perform a pre-processing on the lineage formula and construct the causality matrix. The causality matrix related to the lineage formula is presented in Fig. 4. This representation could help us to get a deep and complete insight on the causes and their probabilities, and help us to compute responsibility and blame so easily. Each column represents a cause with respect to each local lineage formula, double vertical lines refer to conjuncts of the local lineage

W	P						
W1	0.0022	(x = 1,0.4)	(u = 2,0.3)	(y = 2,0.25)	(v = 2,0.25)	(z = 4,0.3)	(w = 2,1)
W2	0.022			(y = 2,0.25)	(v = 2,0.25)	(z = 4,0.3)	(w = 2,1)
W3	0.0033	(x = 2,0.6)	(u = 2,0.3)			(z = 4,0.3)	(w = 2,1)
W4	0.055					(z = 3,0.7)	(w = 2,1)

Fig. 4. Causality matrix.

```

function COMPUTECAUSES( $\lambda_a^{Q,D}$ )
2: for each  $\lambda_a^{Q,W}$  in  $\lambda_a^{Q,D}$  do
    for each minterm  $C$  in  $\lambda_a^{Q,W}$  do
4:        $Causes\_W = Causes\_W \cup \langle V_i = v_i; P(V_i = v_i) \rangle$ 
    end for
6:    $Causes = Causes \cup Causes\_W$ 
    end for
8: end function

function COMPUTERESPONS(Causes)
    for each Distinct cause  $C$  in  $Causes\_W$  do
3:    $K = 0$ 
    for each minterm  $M$  in  $\lambda_a^{Q,W}$  do
        if  $C \notin M$  then
6:            $K = K + 1$ 
        end if
    end for
9:    $dr(C) = 1 / (1 + K)$ 
     $Pr(C) = Pr(W)$ 
     $drP(C) = drP(C) + (drP(C) \times Pr(C))$ 
12: end for
end function

function COMPUTECAUSES(Causes with dr and Pr)
    for each uncertain attribute  $X$  do
        for each cause  $C$  in  $Causes\_W: C \in X_T^W$  do
4:            $dB(X) = \sum_{w \in W_t} \prod_{t \in T_{v_i}^w} P(V_i = v_i) dr_t(V_i = v_i)$ 
        end for
    end for
end function

function OUTPUTEXPLANATIONS(Causes with dR and Pr,
Attributes with dB)
    Output Causes ordered by  $dR \times Pr$ 
3:   Output Attributes ordered by  $dB$ 
end function

```

formula. Computing responsibility is based on computing the number of conjuncts that do not include the cause. Let us take the cause $z = 3$, $z = 4$ and $w = 2$, and compute their responsibilities along the 4 worlds.

$$\begin{aligned}
 dr_t(w = 2)^{w_1} &= dr_t(z = 4)^{w_1} = 1/3 \\
 dr_t(w = 2)^{w_2} &= dr_t(z = 4)^{w_2} = 1/2 \\
 dr_t(w = 2)^{w_3} &= dr_t(z = 4)^{w_3} = 1/2 \\
 dr_t(w = 2)^{w_4} &= dr_t(z = 3)^{w_4} = 1
 \end{aligned}$$

In the first world $dr_t(w = 2)^{w_1} = 1/3$ since there exists two other conjuncts that do not contain $w = 2$, that is, in order to make $w = 2$ counterfactual cause, there exists two other contingencies. In W_2 and W_3 , there exists one contingency, $dr_t(w = 2)^{w_2} = 1/2$. In the world W_4 , $w = 2$ has a degree of responsibility $dr_t(w = 2)^{w_4} = 1$, because it is a counterfactual cause.

Now for computing probabilistic responsibility, we have all information in hand, so the probabilistic responsibilities of these causes are computed as follows:

$$\begin{aligned} drP_t(w=2) &= dr_t(w=2)^{w_1} \times P(W_1) + dr_t(w=2)^{w_2} \times P(W_2) \\ &+ dr_t(w=2)^{w_3} \times P(W_3) + dr_t(w=2)^{w_4} \times P(W_4) \\ &= (1/3) \times 0.0022 + (1/2) \times 0.022 + (1/2) \times 0.0033 + (1) \times 0.055 \\ &= 0.0683 \end{aligned}$$

$$\begin{aligned} drP_t(z=4) &= dr_t(z=4)^{w_1} \times P(W_1) + dr_t(z=4)^{w_2} \times P(W_2) \\ &+ dr_t(z=4)^{w_3} \times P(W_3) \\ &= (1/3) \times 0.0022 + (1/2) \times 0.022 + (1/2) \times 0.0033 \\ &= 0.013 \end{aligned}$$

$$drP_t(z=3) = dr_t(z=3)^{w_4} \times P(W_4) = 1 \times 0.055 = 0.055$$

We notice that $drP_1(z=3) > drP_1(z=4)$ although $z=4$ is included in more worlds than $z=3$ (W_1, W_2, W_3). However, since $z=3$ is a counterfactual cause in a world with a high probability, that makes it more responsible cause. It is evident here that the most responsible cause is $w=2$, since it is included in all worlds.

Now let us compute the blame for each uncertain attribute SSN and M. We recall that variables x, y and z are defined under the attribute SSN, whereas u, v and w are defined under the variable M.

$$\begin{aligned} db(SSN) &= [(dr_t(x=1)^{w_1} \times P(x=1)) \times (dr_t(y=3)^{w_1} \times P(y=3))] \\ &\times (dr_t(z=4)^{w_1} \times P(z=4))] \\ &+ [(dr_t(y=1)^{w_2} \times P(y=1)) \times (dr_t(z=4)^{w_2} \times P(z=4))] \\ &+ [(dr_t(x=2)^{w_3} \times P(x=2)) \times (dr_t(z=4)^{w_3} \times P(z=4))] \\ &+ [(dr_t(z=4)^{w_4} \times P(z=4))] \\ &= (0.4 \times 0.3 \times 0.3 \times 1/3) + (0.7 \times 0.3 \times 1/2) + (0.6 \times 0.3 \times 1/2) + \\ &0.7 \times 1 = 0.907 \end{aligned}$$

$$\begin{aligned} db(M) &= [(dr_t(u=2)^{w_1} \times P(u=2)) \times (dr_t(v=2)^{w_1} \times P(v=2))] \\ &\times (dr_t(w=2)^{w_1} \times P(w=4))] \\ &+ [(dr_t(v=2)^{w_2} \times P(v=2)) \times (dr_t(w=2)^{w_2} \times P(w=2))] \\ &+ [(dr_t(u=2)^{w_3} \times P(u=2)) \times (dr_t(w=4)^{w_3} \times P(w=2))] \\ &+ [(dr_t(w=2)^{w_4} \times P(w=2))] \\ &= 0.3 \times 0.25 \times 1 \times 1/3 + 0.25 \times 1 \times 1/2 + 1 \times 1/2 + 1 \times 1 \\ &= 1.314 \end{aligned}$$

We see here that $db(M) > db(SSN)$, that is M contributes the most for the probability of the output result of SSNs for our query. One evident difference between the two attributes is the certain tuple with the valuation $w=1$ and responsibility one in W_4 . This shows clearly that the attribute having the most blame is the one consisting of more responsible and certain tuples.

We should mention that this measure informs us clearly about which attribute contributes more to the uncertainty that impacts the output probability. Let us suppose the case where SSNs are usually certain, and just under some cases the probability is distributed over two possible values at most, and for M, we suppose that along all forms the value of M is never certain and might take all 4 possible values. Considering a case like this would result evidently in $db(SSN) > db(M)$, i.e, SSN contributes the most for the output probability. Actually, knowing such information will have a high importance through enabling us to address the source of this uncertainty in the future.

V. EXPERIMENTAL EVALUATION

We implemented the above algorithm in C#. We tested our method in Windows 10 with i7 CPU and 8 GB memory. To our knowledge, there is no benchmark of probabilistic databases available for performing experiments, therefore, we have dealt with this issue through two

main experiments on synthetic and real data respectively. In the first experiment, we created a probabilistic database that concerns the census scenario by creating a list of random forms, and then we tried to compute causes and responsibilities. This experiment is meant mainly to measure the performance of our method in terms of execution time. The second experiment is based on a real data set extracted from the IMDB database [24]. In this experiment we will show the usefulness of causality, responsibility and blame for explaining probabilistic query answers.

A. Synthetic Census Data Set

We make two basic experiments on the synthetic census data set. The first is done by fixing the number of worlds by 500 and varying the number of forms. We compute the time execution in seconds for different numbers of forms, from 5000 to 30000 forms.

The result of this experiment is depicted in Fig. 5. For 5000 forms, the time execution of our algorithm is estimated by 0.79 seconds, and 1.36 seconds for 10000, and continues increasing, until it reaches 10.57 seconds for 30000 forms.

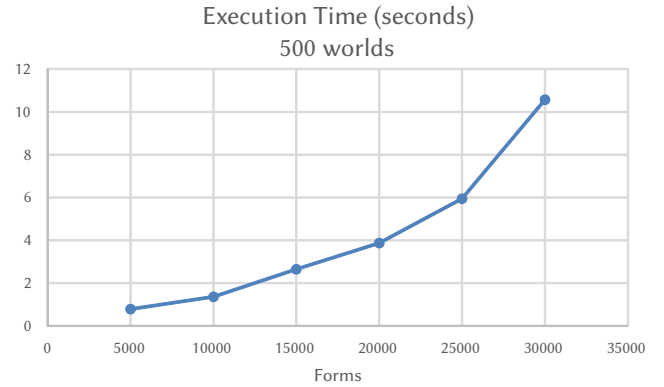


Fig. 5. Execution time with respect to number of forms.

We perform another experiment now by fixing the number of forms by 50000, and varying the number of worlds from 100 to 600 (see Fig. 6). For 100 worlds, the time taken by our algorithm is estimated by 1.36 seconds, and 3.71 for 200 worlds until it reaches 19.60 seconds for 600 worlds. As a result of these two experiments, we see here that the execution time is affected by both the number of forms and worlds in similar way, and the computation time is efficient even for large sizes. In general, we observe that computing explanations is linear in the size of the number of forms as well as as the number of worlds. In other words it is linear in the size of the lineage, which conforms with the results found by Kanagal and Deshpande [13], where the computation times are also very similar.

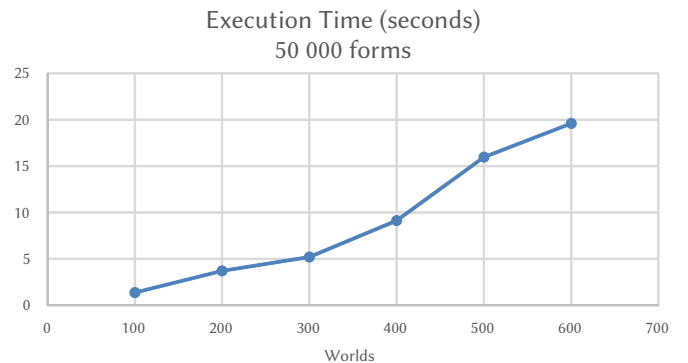


Fig. 6. Execution time with respect to number forms given 100 worlds.

Now we are going to present the results of the number of probabilistic responsibility classes based on the same previous experiments. So, we first fix the number of worlds by 500 and change the number of forms, and then fix the number of forms by 50000 and change the number of worlds. The results are depicted in Fig. 7 and Fig. 8 respectively. As we see in Fig. 7, it is evident that the number of causes and their probabilistic responsibilities increases in function of the number of forms. The second observation, which is more important, is that we have a small number of classes comparing to the number of causes. For instance, for 5000 forms, we have 2506002 causes and just 455 classes of probabilistic responsibilities, and given 50 000 forms, we have over 17 million (17536002) causes with 501 classes only. We can explain this result by having a large set of causes that share the same probabilistic responsibility value. Regrouping together all the causes that share the same probabilistic responsibility would be very useful. However, we notice that the number of classes is close from a value to another, where is still constant starting from 20000 forms. This is can be explained as the following: as long as the number of causes increases, we will have equal probabilistic responsibilities values, and thus no new classes are created. We should mention here that the number of causes does not refer to distinct causes, but rather to the number of causes counted over all the worlds, which means that the causes counted in a world are counted again in another world, which is required, since we are interested in computing probabilistic responsibilities at each world.

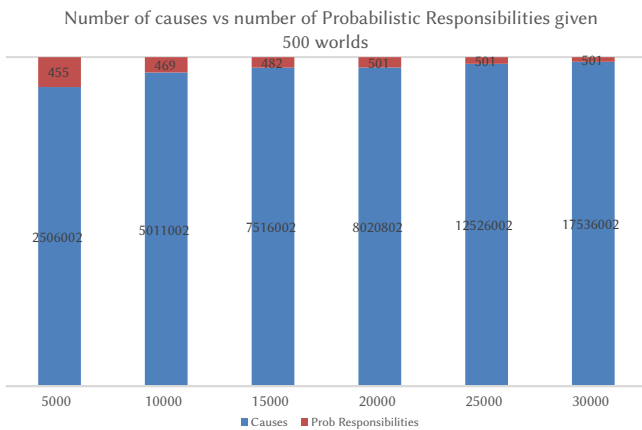


Fig. 7. Causes and probabilistic responsibility in function of forms number.

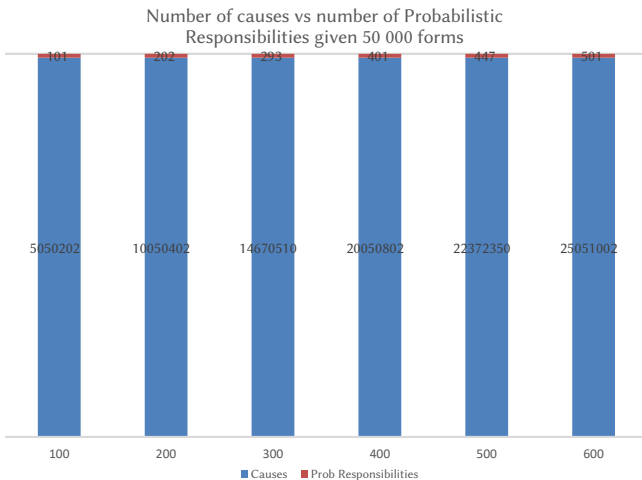


Fig. 8. Causes and probabilistic responsibility in function of worlds number.

Concerning the second experiment, we see that the results depicted in Fig. 8 are similar to the previous results. It is evident that as the number of worlds increases, we obtain a large set of causes as well as

probabilistic responsibility classes. We notice that the maximum value of classes number, which is 501 is reached at 600 worlds, which refers exactly to the same maximum value found in the previous experiment.

Now we are going to present a final experiment that aims to show the impact of fixing the number of probabilistic responsibility classes on speeding up the execution time. Fixing the maximum value for probabilistic responsibilities is desirable by the user in order to focus only on most responsible causes and omitting the causes of low importance, which is one of the objectives of our work. To better show the impact of fixing the number of classes, we depict in Fig. 9 the execution time for top 100 classes together with the results of the first experiment. As we see in the results, the execution time is remarkably reduced. For instance for 5000 forms, the time is reduced from 0.79 to 0.24, from 1.36 to 0.38 for 10000 forms, and finally for 30000 forms, the time is reduced from 10.57 seconds to 1.14 seconds. So, relying on most responsible causes only allows to focus on important causes, and thus it helps to deliver good results in term of execution time, since we are omitting a large set of causes that share the rest of probabilistic responsibility classes, which are estimated for instance for 30000 forms by 400 classes.

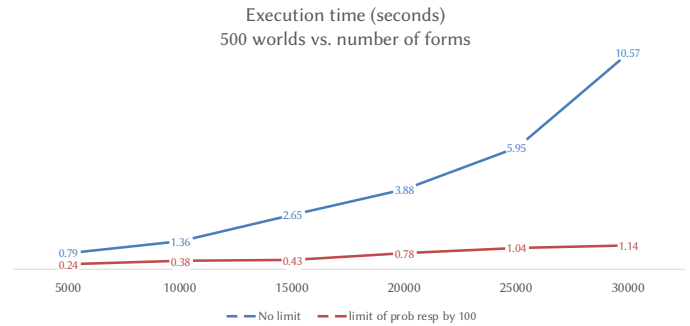


Fig. 9. Difference in execution time for 100 classes of probabilistic responsibility.

B. IMDB Data Set

This experiment is based on a prepared probabilistic database that has uncertain relations, where the sources of uncertainty are related to fuzzy object matches for titles, and the confidence in movies ratings [46]. The probabilistic IMDB database consists of two main relations. The schema of this database is presented in Fig. 10. While the first relation introduces uncertainty on the level of title, the second introduces uncertainty on the level of ratings, where a rating has a value in the range: 1 to 5.

```

Movies(movie_id int, title varchar,
year int, director varchar)
Ratings(movie_id int, cust_id int,
date varchar, rating int, confidence float)
    
```

Fig. 10. IMDB probabilistic database schema.

The movies relation consists of 1881 tuples, and the ratings relation consists of 10037 tuples. We aim to execute a query that returns the possible ratings for such a movie. All the queries required for returning this result are introduced in the database management system MayBMS [22]. Based on the possible worlds semantics, we must first repair the database to enforce constraints and make it consistent through the following queries:

```

create table ratings_1 as
select rating, movie_id
from (repair key movie_id in
ratings weight by confidence) r;
    
```

```

-----
create table movies_1 as
select movie_id, title, director, year
from (repair key movie_id in movies) q;

```

Now, we estimate the exact confidence of distinct tuples through *conf()* function, such as: what are the possible ratings for each movie given all the instances of the probabilistic database. *Conf()* is computed as the sum of the probabilities of the instances (worlds) in which the tuple occurs as follows:

```

create table ratings_1_conf as
select rating, movie_id, conf()
from ratings_1
group by movie_id, rating;
-----
create table movies_1_conf as
select title, movie_id, director,
year, conf()
from movies_1
group by movie_id, title, director, year;

```

Now, we can join the two relations into a third one to obtain complete information using the following query:

```

create table ratings_movies as
select M.movie_id, M.title,
M.year, R.rating, R.conf
from movies_1_conf as M,
ratings_1_conf as R where
M.movie_id=R.movie_id
group by M.movie_id, M.title,
M.year, M.director, R.rating,
R.conf order by M.movie_id,

```

The resulting relation *ratings_movies* consists of 3292 tuples. Given this relation, we now want to get the possible ratings of such a movie. We choose *Jack* as a title for our query. The query as well as its results is given below:

```

select distinct rating, conf from (select *
from ratings_movies_conf where title
like '%Jack%') Q;

```

```

-----
rating      | conf
-----
1 | 0.12
2 | 0.02
2 | 0.2
2 | 0.3
3 | 0.42
3 | 0.52
3 | 0.56
4 | 0.06
4 | 0.12
4 | 0.32
5 | 0.12
5 | 0.24

```

As we see in the results, the user can get different confidence values for different ratings values, where all ratings values are present from 1 to 5. The user can have in mind different questions regarding these results, and he would be seeking for some explanations. Why I have rating 1 in my results, while I'm expecting at least 3? why 3 has the highest probabilities values, while I'm expecting the movie to have

rating 5? is there a mismatch with other movies? so who is the director of this movie, and in which year? etc. In other words, which tuples are responsible for such an outcome.

The tuples responsible for the presented output are 248 tuples, which can be returned by the following query:

```

$select * from ratings_movies_conf
where title like '%Jack%'$.

```

We see here that for a specific requested movie (*Jack*), we have 248 tuples that have contributed, which clearly shows that the evaluation of a lineage formula, and extracting the most responsible causes for a query answer in probabilistic databases is a challenging task. For each resulting rating value, we investigate the lineage information and the most responsible causes. For the value 1, there is one tuple that has contributed, which is a conjunction of two tuples from *movies* and *ratings* relations. These tuples are: *movies(581, BlackJack: Themovie, 1993, Unknown)* and *ratings(581, 1)*, which are returned as an explanation for the rating 1. Both tuples have a degree of responsibility 1, since both tuples are counterfactual causes, where removing one tuple no longer yields a rating 1. However, the tuple *movies(581, BlackJack: Themovie, 1993, Unknown)* is the most responsible cause, since it is certain, thus it has a degree of probabilistic responsibility (*drP*) greater than the tuple *ratings(581, 1)*, which is not certain. This example clearly shows the need for probabilistic responsibility, where classical responsibility is not enough for explaining probabilistic databases.

In contrast to rating 1, for the values 2, 3, and 4, we have 62 tuples contributing to the query answers, whereas for rating 5, we have 61 tuples, which results in 248 tuples in total. An example of results for rating 2, rating 3 and rating 4 are presented in Fig. 11. For all of these values, we see that we have exactly two tuples of the movies 508, and 581, whereas for the rest of 60 tuples, they are all related to the movie 172. This means that the movies 508 and 581 are certain tuples, whereas the movie 172 introduces a high level of uncertainty with different matches on titles. For instance, we have the same title *Jack* in different years (1913, 1916,...) and different titles matches (*Jack, Jacky, Jill Jacks Off, ...*). Now we want to compute the probabilistic responsibility for both movies and ratings tuples.

Since the movies 508 and 581 are certain, they appear in all the worlds of this probabilistic database, and since in every world we have 3 movies with the title *Jack*, the responsibility of these causes is 1/3, and thus for rating 2, the most responsible cause is the tuple: *movies(581, Black Jack: The movie, 1993, Unknown)*, since it has the highest degree of probabilistic responsibility (*drP*). For rating 3, the confidence seems high and close for every tuple, which means that it is the highly possible rating for all movies. Although, the Sixteen (60) movies with the *movie_id = 172* seem to have the highest confidence values, every tuple can occur only once at each world, in contrast to the two previous ones (508 and 581), which are certain, and therefore the most responsible cause is still the same (*t:movies(581, Black Jack: The movie, 1993, Unknown)*). The results for rating 4 are very similar, however, the most responsible cause this time is *movies(508, Siant Jack, 1979, Bogdanovich Peter)*. For rating 5, we have only two movies that are candidate to be causes, the movie 172 and 508, where 581 has no responsibility at all. The most responsible cause is *movies(508, Siant Jack, 1979, Bogdanovich Peter)*, since it is a certain tuple, and its probability of being ranked 5 is higher than the movie 172. To clarify more the importance of these results for explanation, let us consider the following scenario: when the user asks for the rating of *Jack*, he was anticipating at least rating 4, which makes him surprised getting ratings: 1, 2 and 3. From the previous results, he would know that the movie is looking for it might be *movies(508, Siant Jack, 1979, Bogdanovich Peter)*, since it has no possibility of being ranked 1, ranked 2 and 3 with low probabilities, and ranked 4 and 5 with high probabilities (most responsible).

	Query results for rating 2	Query results for rating 3	Query results for rating 4
1	508 Saint Jack 1979 Bogdanovich Peter 2 0.02	508 Saint Jack 1979 Bogdanovich Peter 3 0.42	508 Saint Jack 1979 Bogdanovich Peter 4 0.32
2	581 Black Jack: The Movie 1993 Unknown 2 0.3	581 Black Jack: The Movie 1993 Unknown 3 0.52	581 Black Jack: The Movie 1993 Unknown 4 0.06
3	172 Jack 1913 Liabel Andre 2 0.2	172 Jack 1913 Liabel Andre 3 0.56	172 Jack 1913 Liabel Andre 4 0.12
4	172 Jack 1916 Borzage Frank 2 0.2	172 Jack 1916 Borzage Frank 3 0.56	172 Jack 1916 Borzage Frank 4 0.12
5	172 Jacky 2000 Hu Fow Pyng 2 0.2	172 Jacky 2000 Hu Fow Pyng 3 0.56	172 Jacky 2000 Hu Fow Pyng 4 0.12
.....
62	172 Jill Jacks Off 1993 Constantinou, S.D 2 0.2	172 Jill Jacks Off 1993 Constantinou, S.D 3 0.56	172 Jill Jacks Off 1993 Constantinou, S.D 4 0.12

Fig. 11. Query results for ratings 2–4.

Now concerning blame, we want to measure the degree of blame for each uncertain attribute, which are *title* and *rating*. Actually, the degrees of blame of the previous query are very close, however, for most values of ratings, the degree of blame of *title* is higher than the degree of blame for *rating* ($db(title) > db(rating)$), which means that it has more contribution for the estimated probabilities, and thus *rating* has the highest source of uncertainty. The results can be explained as the following: as we showed before, as much as the uncertain relation has certain tuples, as much as the degree of blame(*db*) increases. Actually, it is the case here for the attribute *title* with the two movies 508 and 581, which are certain, however for the movie 172, we have a high level of uncertainty with sixty (60) possible titles match, which decreases significantly the degree of blame for this uncertain attribute. Though, the degree of blame for *rating* is low comparing to *title*, because all the movies concerned with this query have no certain rating values. Now, if we suppose that the movie 172 is certain as well, that will result absolutely in increasing significantly the degree of blame of *title*. We should mention that this kind of information is not just helpful for estimating the level of uncertainty for each uncertain attribute on the local level of a query answer, but also it can give us a better insight on the entire probabilistic database and how these attributes affect our queries results.

VI. CONCLUSION AND FUTURE WORK

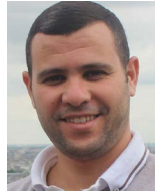
In this paper, we have shown how causality, responsibility and blame can be used in a complementary way in order to give useful quantitative explanations for query results in probabilistic databases. We presented how probabilistic responsibility measure could help us to identify which tuples have contributed the most for a query answer. In addition, we employed blame to identify among many uncertain attributes, which attribute has the most blame for such an outcome. This technique enables us to obtain a complete explanation framework. This technique has been proved to be useful in probabilistic database management systems that feature U-relational model like MayBMS. We delivered an algorithm for computing explanations, which has been implemented and tested on synthetic as well as real data sets.

While our technique addresses only conjunctive queries, among future works is the study of other types of complex queries, such as aggregation and top-k queries. Furthermore, this framework may benefit from employing sufficient lineage instead of complete lineage, which produces a smaller approximate lineage formula.

REFERENCES

- [1] Carnegie Mellon University, "Read the Web" Research Project Website Accessed: Oct. 19, 2022. [online]. available: <http://rtw.ml.cmu.edu/rtw/>.
- [2] D. Suciu, D. Olteanu, C. Re, C. Koch, *Probabilistic Databases*. Morgan and Claypool Publishers, 2010.
- [3] P. Bosc, O. Pivert, "Modeling and querying uncertain relational databases: A survey of approaches based on the possible worlds semantics," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 18, no. 5, pp. 565–603, 2010.
- [4] N. Dalvi, D. Suciu, "Efficient query evaluation on probabilistic databases," *The International Journal on Very Large Data Bases*, vol. 16, no. 4, pp. 523–544, 2007.
- [5] X. Dong, Y. Luming, "Study on consistent query answering in inconsistent databases," *Frontiers of Computer Science in China*, vol. 1, no. 4, pp. 493–501, 2007.
- [6] M. Arenas, L. E. Bertossi, J. Chomicki, "Consistent query answers in inconsistent databases," in *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 1999, pp. 68–79.
- [7] F. Parisi, J. Grant, "On repairing and querying inconsistent probabilistic spatio-temporal databases," *International Journal of Approximate Reasoning*, vol. 84, pp. 41–74, 2017.
- [8] M. Calautti, L. Libkin, A. Pieris, "An operational approach to consistent query answering," in *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 2018, pp. 239–251.
- [9] X. Wang, X. L. Dong, A. Meliou, "Data x-ray: A diagnostic tool for data errors," in *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 2015, pp. 1231–1245.
- [10] C. Berkholtz, M. Merz, "Probabilistic databases under updates: Boolean query evaluation and ranked enumeration," in *Proceedings of the 40th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 2021, p. 402–415.
- [11] L. Antova, T. Jansen, C. Koch, D. Olteanu, "Fast and simple relational processing of uncertain data," in *In Proc. 24th IEEE International Conference on Data Engineering*, 2008, pp. 983–992.
- [12] G. V. den Broeck, D. Suciu, "Query processing on probabilistic data: A survey," *Foundations and Trends in Databases*, vol. 7, no. 4, pp. 197–341, 2015.
- [13] B. Kanagal, J. Li, A. Deshpande, "Sensitivity analysis and explanations for robust query evaluation in probabilistic databases," in *ACM SIGMOD International Conference on Management of Data*, 2011, pp. 841–852.
- [14] C. Re, D. Suciu, "Approximate lineage for probabilistic databases," *The International Journal on Very Large Data Bases*, vol. 1, no. 1, pp. 797–808, 2008.
- [15] I. Ceylan, S. Borgwardt, T. Lukasiewicz, "Most probable explanations for probabilistic database queries," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, 2017, pp. 950–956.
- [16] J. Y. Halpern, J. Pearl, "Causes and explanations: A structural-model approach part i: Causes," in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, 2001, pp. 194–202.
- [17] J. Y. Halpern, J. Pearl, "Causes and explanations: A structural-model approach. part ii: Explanations," *British Journal for the Philosophy of Science*, vol. 56, no. 4, pp. 889–911, 2008.
- [18] H. Chockler, J. Y. Halpern, "Responsibility and blame: a structural model approach," *Journal of Artificial Intelligence Research (JAIR)*, vol. 22, no. 1, pp. 93–115, 2004.
- [19] A. Meliou, W. Gatterbauer, J. Y. Halpern, C. Koch, "Causality in databases," *IEEE Data Engineering Bulletin*, vol. 33, no. 3, pp. 59–67, 2010.
- [20] L. Bertossi, B. Salimi, "Causes for query answers from databases: Datalog abduction, view-updates, and integrity constraints," *International Journal of Approximate Reasoning*, vol. 90, pp. 226–252, 2017.
- [21] L. Antova, T. Jansen, C. Koch, D. Olteanu, "Fast and simple relational processing of uncertain data," in *IEEE 24th International Conference on*

- Data Engineering*, 2008, pp. 983–992.
- [22] L. Antova, C. Koch, D. Olteanu, “Maybms:managing incomplete information with probabilistic world-set decompositions,” in *In Proc. 23rd IEEE International Conference on Data Engineering*, 2007, pp. 1479–1480.
- [23] D. Suciu, “Probabilistic databases for all,” *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 2020, p. 19–31.
- [24] IMDb.com, Inc., IMDb Website. Accessed: Oct. 19, 2022. [Online]. Available: <https://www.imdb.com/>.
- [25] A. Meliou, W. Gatterbauer, K. Moore, D. Suciu, “The complexity of causality and responsibility for query answers and non-answers,” *The International Journal on Very Large Data Bases*, vol. 4, no. 1, pp. 34–45, 2010.
- [26] T. J. Green, V. Tannen, “Models for incomplete and probabilistic information,” in *Proceedings of the international conference on Current Trends in Database Technology*, 2006, pp. 278–296.
- [27] T. J. Green, G. Karvounarakis, V. Tannen, “Provenance semirings,” in *Proceedings of the 2007 ACM SIGMOD- SIGACT-SIGAI Symposium on Principles of Database Systems*, 2007, pp. 31–40.
- [28] R. Diestelkämper, S. Lee, M. Herschel, B. Glavic, *To Not Miss the Forest for the Trees - A Holistic Approach for Explaining Missing Answers over Nested Data*, p. 405–417. 2021.
- [29] B. Salimi, *Quer-Answer Causality in Dataabases And its Connections with Reverse Reasoning Tasks in Data And Knowledge Management*. PhD dissertation, Carleton University, 2015.
- [30] L. Bertossi, B. Salimi, “From causes for database queries to repairs and model-based diagnosis and back,” *Theory of Computing Systems*, vol. 61, no. 1, pp. 191–232, 2017.
- [31] A. Meliou, A. Meliour, S. Nath, D. Suciu, “Tracing data errors with view-conditioned causality,” in *Proceedings of the 2011 ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 2011, pp. 505–516.
- [32] X. Lian, L. Chen, “Causality and responsibility: probabilistic queries revisited in uncertain databases,” in *Proceedings of the 22nd ACM international conference on Information and Knowledge Management*, 2013, pp. 349–358.
- [33] K. Mu, “Responsibility for inconsistency,” *International Journal of Approximate Reasoning*, vol. 61, pp. 43–60, 2015.
- [34] Z. Miao, Q. Zeng, B. Glavic, S. Roy, “Going beyond provenance: Explaining query answers with pattern-based counterbalances,” in *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD ’19, 2019, p. 485–502.
- [35] L. Antova, C. Koch, D. Olteanu, “From complete to incomplete information and back,” in *ACM SIGMOD International Conference on Management of Data*, 2007, pp. 713–724.
- [36] J. Boulos, N. Dalvi, B. Mandhani, S. Mathur, C. Ré, D. Suciu, “Mystiq: a system for finding more answers by using probabilities,” in *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, 2005, pp. 891–893.
- [37] O. Benjelloun, A. D. Sarma, A. Halevy, J. Widom, “Databases with uncertainty and lineage,” in *The International Journal on Very Large Data Bases*, 2006, pp. 953–964.
- [38] O. Benjelloun, A. D. Sarma, C. Hayworth, J. Widom, “An introduction to uldbs and the trio system,” *IEEE Data Engineering Bulletin*, vol. 29, no. 1, pp. 5–16, 2006.
- [39] T. Imielin’ski, W. Lipski, “Incomplete information in relational databases,” *Journal of the ACM (JACM)*, vol. 31, no. 4, pp. 761–791, 1984.
- [40] P. Sen, A. Deshpande, L. Getoor, “Read-once functions and query evaluation in probabilistic databases,” *The International Journal on Very Large Data Bases*, vol. 3, no. 1, pp. 1068–1079, 2010.
- [41] D. Olteanu, J. Huang, “Using obdds for efficient query evaluation on probabilistic databases,” in *2nd International Conference on Scalable Uncertainty Management*, 2008, pp. 109–121.
- [42] A. Darwiche, P. Marquis, “A knowledge compilation map,” *Journal of Artificial Intelligence Research*, vol. 17, no. 1, pp. 229–264, 2002.
- [43] C. Re, N. Dalvi, D. Suciu, “Efficient top-k query evaluation on probabilistic data,” in *2007 IEEE 23rd International Conference on Data Engineering*, 2007, pp. 108–122.
- [44] C. Koch, D. Olteanu, “Conditioning probabilistic databases,” *The International Journal on Very Large Data Bases*, vol. 1, no. 1, pp. 313–325, 2008.
- [45] A. Meliou, W. Gatterbauer, K. Moore, D. Suciu, “Why so? or why no? functional causality for explaining query answers,” Department of Computer Science and Engineering, University of Washington, Seattle, 2010.
- [46] MayBMS Project, MayBMS - A Probabilistic Database Management System. Accessed: Oct. 19, 2022. [Online]. Available: <http://maybms.sourceforge.net/>.



Hichem Debbi

He received his Master’s and PhD degrees in computer science from the University of M’sila, Algeria in 2011 and 2015 respectively. He is currently an assistant professor at the department computer science, University of M’sila. His research interests include but not limited to: causality, verification and explanation of probabilistic systems, debugging and analysing complex systems.

A Platform for Swimming Pool Detection and Legal Verification Using a Multi-Agent System and Remote Image Sensing

Héctor Sánchez San Blas, Antía Carmona Balea, André Sales Mendes, Luís Augusto Silva*, Gabriel Villarrubia González

Expert Systems and Applications Lab—ESALAB, Faculty of Science, University of Salamanca - Plaza de los Caídos s/n, 37008 Salamanca (Spain)

Received 30 September 2021 | Accepted 21 October 2022 | Published 11 January 2023



ABSTRACT

Spain is the second country in Europe with the most swimming pools. However, the legal literature estimates that 20% of swimming pools are not declared or irregular. The administration has a corps of people who manually analyze satellite or drone images to detect illegal or irregular structures. This method is costly in terms of effort and time, and it is also a method based on the subjectivity of the person carrying it out. This proposal aims to design a platform that allows the automatic detection of irregular pools. Using geographic information tools (GIS) based on orthophotography, combined with advanced machine learning techniques for object detection, allows this work. Furthermore, using a multi-agent architecture allows the system to be modular, with the possibility of the different parts of the system working together, balancing the workload. The proposed system has been validated by testing it in different towns in Spain. The system has shown promising results in performing this task, with an F1-Score of 97.1%.

KEYWORDS

Deep Learning, GIS Detection, Illegal Pools Detection, Pool Aerial Recognition.

DOI: 10.9781/ijimai.2023.01.002

I. INTRODUCTION

CARTOGRAPHY deals with the conception, production, dissemination, and study of maps that have undergone a good evolution in the last decade. Until a few years ago, the processes of cartographic revision and, mainly, those aimed at calculating the fiscal area have been carried out manually. Specifically, these processes required significant investments in airplanes or helicopters, making the processes more expensive. Because of this, municipalities are unable to perform cartography surveys frequently. One of the most current research fields is to investigate technological capabilities for local authorities to perform detailed surveys of the territory of municipalities at a reasonable cost.

One of the most important milestones that cartography allows is refining fiscal data or verifying geographic information that forms the basis for local taxes. One aspect taken into account in an audit process by the municipalities is the size of the plots and the construction of private swimming pools. Spain, for tax purposes, needs to take into account a private swimming pool on a plot of land. However, this process is not easy, as there are many aspects to consider with the mapping, such as location, time of day when we take the image, how close or far away the image is, or obstacles. In addition, the result of this process must be evaluated by a person, a costly task that depends

on the individual's subjectivity, as he or she is the one who must detect the structures built in an area. Depending on the image's characteristics relative to the cartography, these characteristics can lead to errors in determining the existence of such structures.

Knowing the precise location of swimming pools is crucial for tax collection purposes and ecological reasons. It is vital to know the pools with large volumes of water since, in the event of a fire in a nearby area, firefighting teams can make use of it [1]. Another major issue that has caused concern in recent years is controlling and managing limited and indispensable water resources such as drinking water. In particular, the construction of swimming pools in summer impacts the demand for water the municipal supplier needs to prepare. Therefore, it is understandable that the local government asks for an extra contribution from pool owners in a tax. A third eminent problem is mosquito-borne diseases, which affect many people worldwide, mainly in tropical and sub-tropical countries such as Brazil. Pool water in unoccupied homes may not be adequately filtered, and rainwater accumulated along with decaying leaves may not be removed from the pool, providing an ideal habitat for mosquitoes to live and breed [2].

With the proliferation and evolution of UAV, particularly the use of RPAS, the process of pool detection is much faster and more cost-effective than it was a decade ago [3]. However, the numerical quantification of pools from a large amount of data produced during the photographic session is a time-consuming task performed by manual procedures. The different satellite images obtained have different properties, making it challenging to develop different algorithms to detect pools. These images differ in scale, resolution, sensor type,

* Corresponding author.

E-mail address: luisaugustos@usal.es

orientation, quality, and ambient illumination conditions. In addition to these difficulties, buildings may have complicated structures and could be hidden by other buildings or trees. Both structural and deterministic clues must be taken into account when constructing the solution. Up-to-date and accurate data are essential for municipalities. Therefore orthophotography makes it slightly easier to recognize outdoor pools. Among the existing possibilities to solve this problem is the use of satellite imagery in combination with machine learning techniques [4].

The main problem in these processes is to relate the images collected by satellites or drones with a pool detection system and the corresponding verification of the same within the databases of local systems. The solution for the detection problem followed the advances in the literature of machine learning algorithms, precisely Deep Learning was used, which is a class of Machine Learning algorithms. This type of algorithm uses multiple layers to progressively extract features from the input images [5].

This article presents a novel platform that allows the automatic detection of swimming pools by acquiring images from different sources. For this purpose, the system can apply different algorithms to determine the proliferation of illegal swimming pools in a territory by accessing local control databases. In order to make the system scalable, robust, and able to merge the information coming from several neural networks, this case study uses a multi-agent architecture. A multi-agent system makes it possible to build a dynamically reconfigurable platform. It also allows the resources and capabilities of the system to be distributed evenly among the different elements of the system. In this way, problems that often occur in centralized systems, such as bottlenecks or recurring access to critical resources, are eliminated. In addition, the system's efficiency in retrieving, filtering, and coordinating information is improved.

The paper is organized as follows: Section II focuses on an in-depth review of the state-of-the-art literature on Deep Learning algorithms used for object detection in images; Section III conducts a review of works similar to the proposed one related to pool detection and pool legality verification; Section IV describes the architecture of the proposed system; On the other hand, Section V will explain each of the blocks that make up the system; Section VI, will show the case study carried out with the results obtained; Finally, the conclusions obtained are in Section VII.

II. BACKGROUND

Deep Learning is a field that focuses on algorithms based on artificial neural networks. Thanks to deep learning, intelligent document processing (IDP) can combine various AI technologies to classify photos automatically and describe the various elements of images. With their multilevel structures, deep learning models are beneficial for extracting detailed information from input images. Convolutional neural networks can also drastically reduce computational time by taking advantage of the GPU for computation, something that many networks do not use. In the field of object identification in images, two methods stand out: Region Proposal algorithms and regression object detection algorithms.

The first method is to find out in advance the possible locations of the target to detect in the figure. This method can ensure that the highest recovery rate is maintained when fewer windows are selected. Suppose an image is an input and, after a series of convolutions and clustering in the backbone, a feature map of size $M \times N$ is obtained, corresponding to the original image's division into $M \times N$ areas. The center of each area from the original image represents the coordinates of a pixel in this feature map.

Region Proposal Algorithms find whether the k anchor boxes corresponding to each pixel contain a target. The network must learn to classify the anchor boxes as background or foreground. It must calculate regression coefficients to modify the foreground anchor box's position, width, and height. Within these classifiers, we find algorithms such as R-CNN [6], Fast R-CNN [7], Faster R-CNN [8] and MASK-CNN [9].

Of the algorithms mentioned above, Mask R-CNN stands out. This algorithm extends Faster R-CNN and works by adding a branch to predict an object mask in parallel with the existing branch for bounding box recognition. The critical element of Mask R-CNN is pixel-to-pixel alignment, which is the main missing piece in Fast/Faster R-CNN. Mask R-CNN adopts the same two-stage procedure with an identical first stage (which is RPN). In parallel to predicting the class and box offset in the second stage, Mask R-CNN also outputs a binary mask for each RoI. This method is in contrast to most current systems, where classification depends on mask predictions. In addition, Mask R-CNN is simple to implement and train thanks to the faster R-CNN framework, which facilitates a wide range of flexible architecture designs. Moreover, the mask branch only adds a small computational overhead, enabling a fast system and rapid experimentation. Fig. 1 shows a visual example of the segmentation performed by the algorithm.

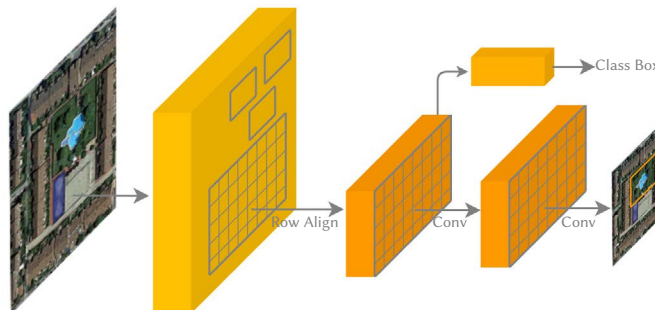


Fig. 1. Mask R-CNN Framework for Instance Segmentation.

Along the same lines as the above, we find Detectron [10], Facebook AI Research's (FAIR) software system that implements state-of-the-art object detection algorithms, including Mask R-CNN. Detectron aims to provide a high-quality, high-performance codebase for object detection research. The design of this algorithm is flexible to support the rapid implementation and evaluation of new research. However, the most successful version is Detectron2 [11], being an enhancement of Detectron. The significant difference between versions is that the latest version is a more modular, flexible, and extensible design, allowing much faster training on GPU-enabled computers. Detectron2 includes high-quality implementations of the most advanced object detection algorithms, such as DensePose, pyramid networks with panoptic features, and numerous variants of the pioneering Mask R-CNN family of models, also developed by FAIR. The creators of the algorithm reproduce the ResNet-50-FPN baselines together with the Scale Jitter algorithm.

A. YoloV4

The above algorithms use detection as a classification problem, that is, first, the algorithm generates object proposals, and then these proposals are sent to the classification/regression regions. However, some methods approach detection as a regression problem based on a similar operation. The YOLO (You Only Look Once) and SSD (Single Shot Detector) algorithms stand out within this field.

The SSD [12] algorithm strikes a good balance between speed and accuracy. SSD runs a convolutional network on the input image only once and computes a feature map. It then runs a small 3×3 convolutional kernel on this feature map to predict bounding boxes

and classification probability. SSD also uses anchor boxes in various aspect ratios, similar to Faster-RCNN, and learns the offset instead of learning the box. To handle scale, SSD predicts bounding boxes after multiple convolutional layers. As each convolutional layer operates at a different scale, it detects objects of various scales. Fig. 2 shows an example of how the SSD algorithm works.

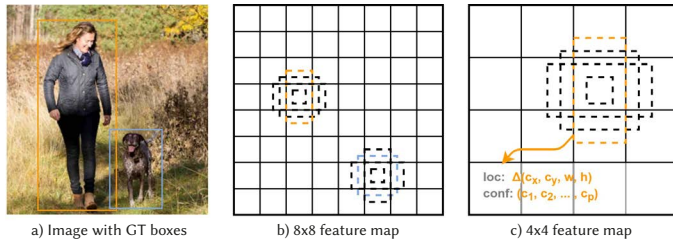


Fig. 2. SSD framework example.

For YOLO [13], detection is a simple regression problem that takes an input image and learns the class probabilities along with the coordinates of the bounding box. YOLO divides each image into an $S \times S$ grid, and each grid predicts N bounding boxes and their confidence. The confidence reflects the accuracy of the bounding box and whether the bounding box contains an object, regardless of the class. YOLO also predicts the classification score of each bounding box for each class in training. It can combine both classes to calculate the probability that each class is present in a predicted box. Thus, the algorithm predicts a total of $S \times S \times N$ bounding boxes. However, most of these boxes have low confidence scores, so if we set a threshold, for example, 30 percent confidence, we can eliminate most of them, as shown in Fig. 3.

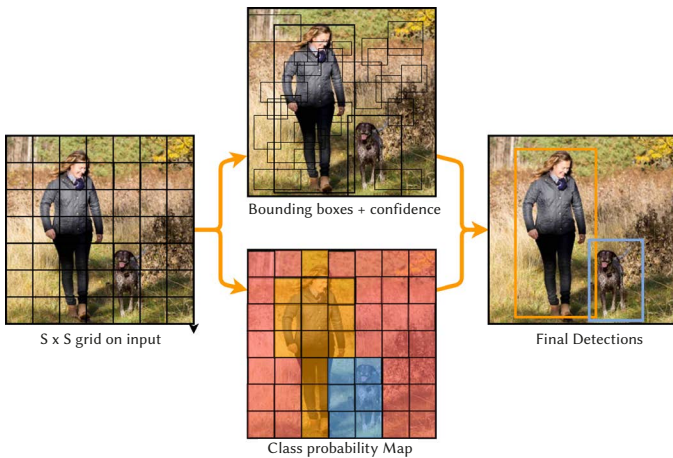


Fig. 3. YOLO workflow example.

YOLO is a algorithm faster than all other detection algorithms, allowing it to run in real-time. Another key difference is that YOLO sees the entire image at once rather than looking only at the proposals of a region generated in previous methods. Thus, this contextual information helps to avoid false positives. However, one of the limitations of YOLO is that it only predicts one class type in a grid, so it has difficulties with tiny objects. There are several versions of YOLO such as YOLOv2 [14], YOLOv3 [15], and YOLOv4 [16]. There are also variations of this latest version, adapting it to the context of use and improving its results [17], [18].

III. RELATED WORKS

The related works listed in this section are recent and present evidence and contributions relevant to the area. However, they are limited and do not have as their primary focus the analyzed points

necessary for controlling and managing pool legality verification based on object detection and image classification.

Thus, we start the comparisons with the work proposed by Tien [1] which introduces a Support Vector Machine (SVM) technique to classify small area water bodies, namely swimming pools. The work focus on locating and subsequently using the water source for emergency services in fighting bushfires in urban areas of Australia. First, satellite images were processed, and then the images were segmented.

In Galindo's [19] work, the authors proposed a detection system to locate full pools during drought periods in order to alert local authorities. The work applies color analysis for water and approximate segmentation and active contouring techniques to refine the shape of the pools. This algorithm with satellite imagery and aerial imagery has a 93% of success rate.

Kim *et al.* [20] shows a concern regarding neglected pools in California, as is the case in other countries [2]. These countries seek economic ease in locating pools for ground survey and mosquito control. This research focused on using high spatial resolution (VHR) satellite imagery and image Pansharping techniques, normalized difference water index, and geographic object-based image analysis. In this way, the authors developed a geographic information system (GIS) database of pool locations. The system demonstrated that VHR imagery could produce a GIS database of pools with high accuracy of 94%.

Rodriguez-Cuenca and Alonso [21] presented in their paper a semi-automatic determination of the location of swimming pools in urban areas from aerial images and LIDAR sensor data. All this without the need for specific training, added indices combined with Dempster-Shafer theory to determine pool locations. The method presented an accuracy rate of 99.86%.

Ferner *et al.* [22] study the detection of homes with swimming pools using convolutional neural networks (CNNs), applied to load heat maps constructed from load profiles. The author presents only a small dataset. Still, the results show that using CNNs, privacy can be broken automatically, without using manual feature generation, which requires a lot of time. The method outperforms the nearest neighbor classifier compared by the authors. Although this work does not use satellite images or aerial images, it is related to the current project as it uses a Convolutional Neural Network (CNN).

Domozi *et al.* [23] also uses a Convolutional Neural Network applied to aerial images by drones. The proposal makes use of an application called Pix4D. The current development allowed to automatically detect the ponds and quantify them employing a neural network on the orthophotos.

Lima *et al.* [24] presented in their article an application integrated into municipal systems, determining the location of swimming pools in urban areas from a GIS system. The application includes the use of deep learning algorithms to detect swimming pools and focus on existing municipal information, with validation in the region of Braga in Portugal. The method has demonstrated an accuracy of 83%, with a significant number of illegal or unknown pools detected.

In this section, we have analyzed different works carried out in the areas involved. To provide a more detailed and clear analysis all the methods explained in this section are summarized in Table I.

IV. PROPOSED ARCHITECTURE

The proposed architecture of this section aims to provide a solution to the problem while adapting and introducing new functionality without affecting the other parts of the system. In order to achieve the objective of this research work, the detection and automatic verification of the legality of swimming pools built in private spaces, it

TABLE I. RELATED WORKS

Work	Year	Algorithm	Images	Accuracy
Tien <i>et al.</i> [1]	2007	Support Vector Machine	Satellite Image	-
Galindo <i>et al.</i> [19]	2009	Color and Segmentation Analysis	Satellite Image	93%
Kim <i>et al.</i> [20]	2011	Pan-sharpening	Satellite Image	94%
Rodríguez-Cuenca [21]	2014	Support Vector Machine	Aerial Image + LiDAR	99.86%
Ferner <i>et al.</i> [22]	2019	CNN / 5-Nearest neighbors	-	68.5% and 71.9%
Domozi <i>et al.</i> [23]	2019	R-CNN	Drone Imagery	99%
Passos <i>et al.</i> [2]	2020	Faster R-CNN / ResNet-101-C4	Drone Imagery	74%
Lima <i>et al.</i> [24]	2021	Faster R-CNN	GeoTIFF Images	83%

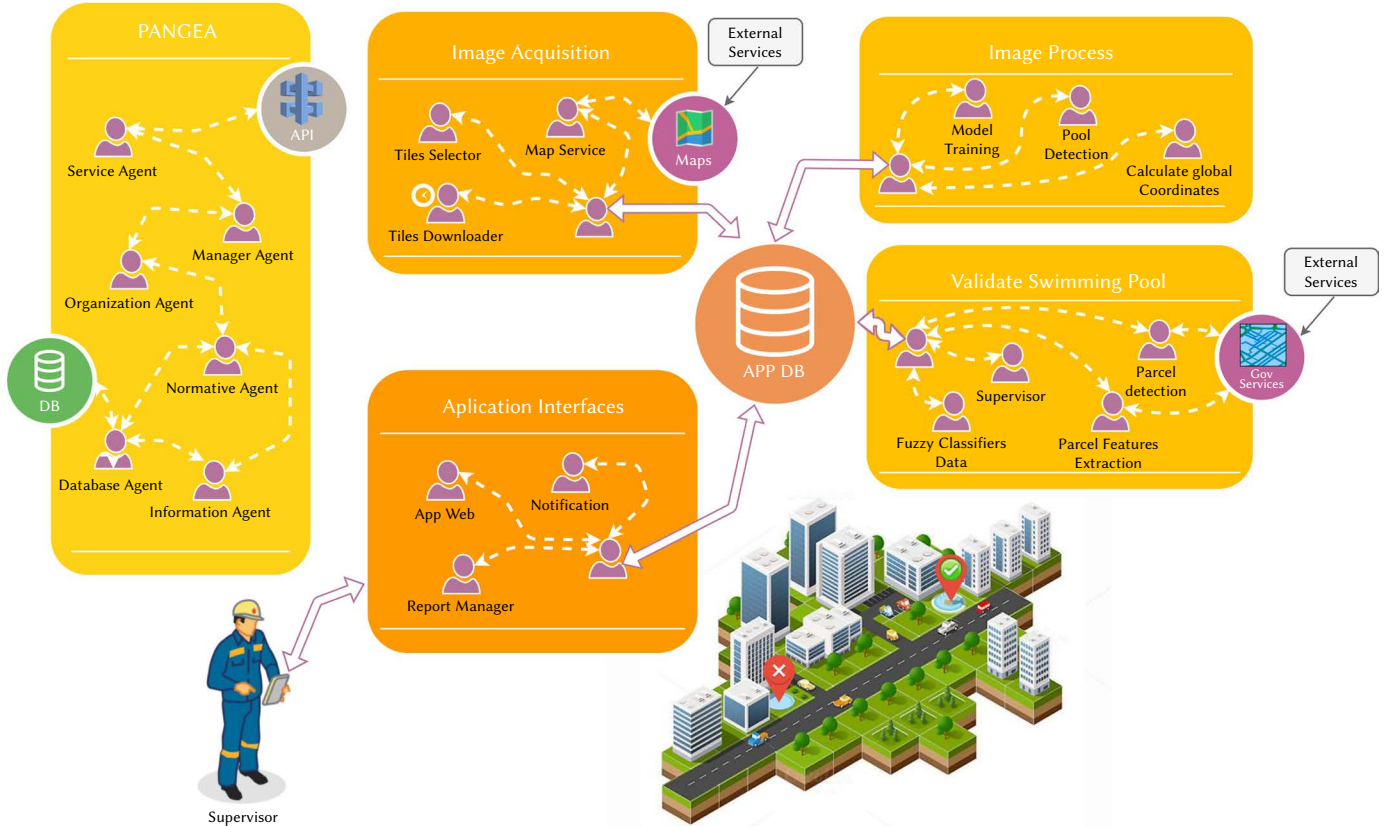


Fig. 4. System Architecture.

is necessary to have an architecture with well-defined characteristics so that the system can operate correctly. For this purpose, an architecture is modeled based on virtual agents, where each of the system’s agents works individually to achieve a common goal. One of the primary needs that have led to the system’s design using this architecture has been the need to design a distributed system to perform the different tasks of image extraction, model training, and classification in a scalable way. Another essential feature of this architecture based on virtual agents is that another can replace one agent without affecting the rest of the system.

PANGEA works [25] have used a base of the proposed architecture. PANGEA allows the different agents to adapt to the computational needs of the system dynamically. Another advantage of this architecture is that it allows the system’s services to rise on demand. When an agent joins the platform, it must communicate which services are available and can offer to other entities. In Fig. 4, the architecture designed for the case study presented in this article is shown. In addition, the agents organized in virtual organizations that make up the system can be seen.

Compared to other existing systems such as SPADE, Python’s Library, JADE, or osBrain, the PANGEA multi-agent platform can

create virtual organizations. These virtual organizations allow the creation of visual representations and modeling of any system. In addition, being an open-source system, it will be available for use by other researchers who want to replicate the system in the future.

The following is a list of the organizations, the functionality, and the agents of each one:

Image Acquisition: This virtual organization is responsible for obtaining the satellite images that the other organizations will use to detect the pools. One of the main features of this organization is that it allows the use of different map sources for downloading the tiles. For this purpose, the Tiles Selector agent is in charge of calculating the tiles to be downloaded for the zones pre-selected by the system users. Analyzing system needs was determined to perform periodic downloads and checks to detect new pools and pools installed temporarily in the summer. For this purpose, the Tiles Downloader agent can program the downloads according to the indicated periodicity.

Application Interface: This organization is the one that allows the information generated by the system to be understandable by humans. This organization serves as an interface between the system and the system’s applications. Applications that have access to this organization can access or generate data in the system. In this case, the

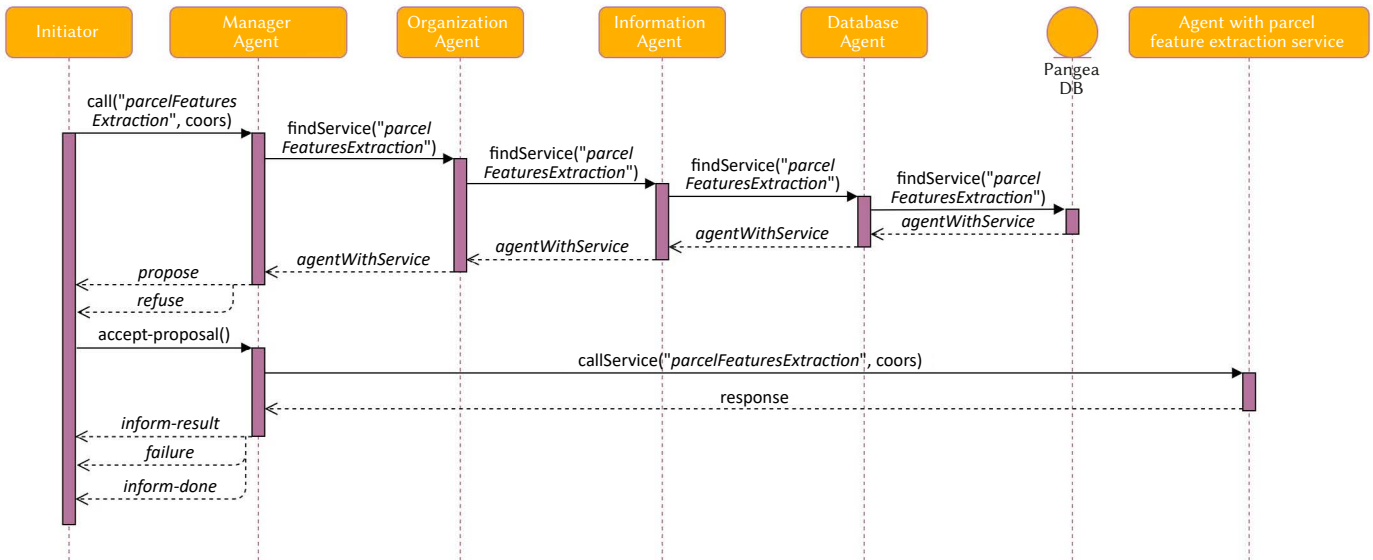


Fig. 5. Sequence Diagram from the System.

human agent with the system can define which zones to inspect and review detection results, alerts, or detection reports.

Image Process: This organization aims to carry out tasks related to image processing and tile pool detection. For this purpose, it has agents with the Model Training agent responsible for the training and retraining the models used with new images labeled by the user. The Pool Detection agent can detect the new tiles pending classification downloaded into the system and proceed to their classification using the pre-trained models. Finally, in this organization, the Calculate Global Coordinates agent's main task is to convert the coordinates of each of the pools detected by the Pool Detection agent to the global system, taking into account the relative coordinates and the zoom level of the tile analyzed.

Validate Swimming Pool: This organization aims to detect which pools are legally registered. To do so, this organization uses a governmental Webservice to obtain the data associated with each of the plots. The information returned by the service is whether any swimming pool is registered. The Parcel Detection agent, using the external services, has the objective of assigning the identifier of the parcel to each of the detections. The parcel uses this identifier Features Extraction agent, which is in charge of obtaining if the pool is registered as a legal form in the parcel. In this way, the system can determine if the detected pool is correctly registered. The Fuzzy Classifiers Data agent is responsible for detecting duplicate pools in adjacent tiles or at a minimal distance.

Pangea Multi-Agent System Organization: In this organization there are the minimum agents necessary for the Pangea system to work. The main milestone of this organization is to carry out the tasks of organizing the virtual organizations and the communication between the agents responsible for each organization. Below the agents that are part of this organization are described:

- **Service Agent:** This agent can expose functionality through web services such as a communication interface between the organization's external agents and those of itself. This interface allows the creation of agents independent of programming language or execution environment.
- **Manager Agent:** Responsible for periodically checking the status of the system, detecting system overloads and possible failures that may occur in the agents of each of the organizations.
- **Organization Agent:** This agent is responsible for verifying the operations of virtual organizations, ensuring security and load

balancing. This agent also provides encryption services.

- **Normative Agent:** It is responsible for enforcing compliance with the rules in communications between agents.
- **Database Agent:** This agent is the only agent in the organization that has database access permissions. It is in charge of storing the system status information, analyzing the data persistence and consistency capabilities.
- **Information Agent:** Responsible for managing the services available within the virtual organizations, indicating which services are available for each of the agents. When an agent joins the system, it must indicate which services are available. In this way, when another agent requests a service, it must query this agent to know which entity is in charge of offering it.

For its correct operation and scalability, the system uses different databases; The system uses the database inside the PANGEA organization to store the information of the system agents, the services provided by each one, and the tasks that any agent can carry out. Apart from this, the system has an additional database used to store the specific information of the case study, the areas selected for inspection, detected pools generated tiles.

One of the advantages offered by this architecture is the contact network, where an external agent can search for and execute a service. To do so, the external agent must send a message to the Manager Agent, indicating the required service with the necessary parameters for that service. In collaboration with the other agents of the PANGEA organization, organization Agent, Information Agent, and Database Agent, this agent responds with a list of available agents that can carry out the requested service. Finally, the agent who is to perform the task must accept the proposal to carry out the task. Fig. 5 shows an example of a request for the service of extracting features from a parcel by an external agent.

V. PROPOSED SYSTEM

The main challenge of this work is to design a platform capable of automatically detecting illegal pools at the lowest possible cost. This process requires data sources that are updated frequently and at the lowest possible cost. Following the current method for this verification, administrations use images of the exterior of people's residences to check existing constructions and verify that they are in the official register. Therefore, the images fulfill the role of a low-cost and up-to-

date data source thanks to GIS tools and represent a method already used by public administrations. The system adapts to the current way of working without creating any problems.

Considering that the primary source of data obtained is images, reviewing the literature, the systems that obtain the best results in classifying and detecting objects in images are Deep Learning algorithms. The use of these algorithms is crucial in developing this task as they will allow the detection of areas where pools exist from the images. Without such a detection capability, the proposed automatic system would not be possible. This section presents the case study based on the sub-block image classification technique. Fig. 6A shows an example of a tile with zoom 18, and Fig. 6B shows an example with zoom 19.

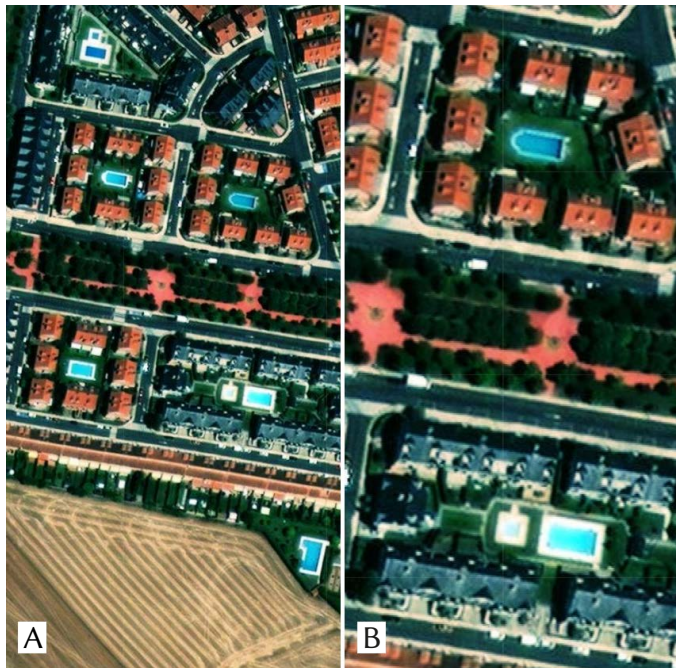


Fig. 6. Satellite Image for Detection with Zoom 18.

The proposed solution has three main blocks, the block of satellite image generation, the block of detection and classification of pools from the generated images, and, finally, the block of checking their legality with municipal databases. In addition, for interconnecting the parts, the PANGEA multi-agent architecture is used, briefly explained in Section IV.

The following sections describe the development steps. Section V.A deals with the development of the image search system in the maps. Section V.B is where the algorithms and classification methods used are presented, contemplating the second block of the work. Added to that is Section V.B.1, which presents the dataset formed by the images and their annotations, followed by the algorithms in sections V.B.2, V.B.3 and V.B.4. Finally, Section V.C describes the system that allows checking if a pool is legally registered.

A. Images Generation System

It is necessary to have a database covering a large area and containing aerial photos to detect pools. Users can build such a database with private or public tools and use expensive large aircraft systems for less expensive drone-based solutions.

The proposed block for the generation of images is interoperable and obtains data from several providers, such as Bing Maps [26], Google Maps [27], OpenStreetMaps [28], ESRI World Imagery [29], Wikimedia Maps [30], NASA GIB S [31], Carto Light [32], Stamen Toner B & W

[33] and the Sentinel [34]. This portfolio of service providers covers a large part of the inhabited areas and are publicly available.

In the present research work, the user draws a zone on the map to inspect the area. In this tool, it is possible to configure, on the one hand, the zoom of the images, and on the other hand, the map data source, as shown in Fig. 7a.

Then, the system generates a grid with small map fragments (tiles) filling the entire drawn area. This process is repeated for each selected zoom as many times as necessary. Fig. 7b illustrates the process of transforming the selected area into the corresponding tiles.

B. Swimming Pool Detection Algorithms

The algorithms used have in common the preprocessing phases of each of the images. The whole process is performed in an identical way and with the same parameters. It is possible to emphasize the process of transformation of the image to grayscale and the subdivision of the image in N blocks of equal size. Subsequently, for each of these blocks, image processing is performed to extract the texture descriptors.

The texture descriptors refer to information about the spatial arrangement of color or intensities in an image. The feature vectors, formed from the texture descriptors, are normalized and used in training the classifiers. The programming language used for the development of the algorithms and experiments was Python 3. In particular, the OpenCV, *skimage* and *mahotas* image processing libraries were used. In addition, the machine learning libraries *scikit-learn*, *TensorFlow*, *keras* and *imbalanced-learn* have been used.

1. Datasets

The dataset used for training is a proprietary dataset. The Images Generation System explained in section V.A has been used to build it. To train the model, 999 images of 512x512 pixels with zoom 18 were obtained from Redlands, CA, around Prospect Park because it has a high concentration of swimming pools. Subsequently, the Roboflow web application [35] was used, which allows for easy labeling of the images. The set of images created consists of a single class, Swimming Pools.

The complete dataset have 778 annotated images with pool and 221 images unannotated, that do not contain any pool. In total, there are 2300 labels of the pool class. All images are randomly arranged in the dataset that was separated in two sets (Table II): the training set (80%) with 799 images and 1892 labels, and the validation set (20%) with 200 images and 408 labels.

TABLE II. NUMBER OF POOLS IN EACH SET OF TRAINING IMAGES

Set	Images	Pools
Training	799	1892
Validation	200	408
All	999	2300

The storage format of the bounding box information for the annotations or labeling of this dataset is different for each object detection model. We need three different storage versions of the bounding box data, one for each object detection model used for the comparison. In [36] the three different datasets are published, one for each detection model, for further verification of the results or evaluation of new algorithms. Fig. 8 shows an example of labeling set.

2. YOLOv4

For the training of the YOLOv4 neural network, we use the Jetson Xavier AGX hardware device. The configuration file used can be found at [37]. It uses the YOLO Darknet annotation format in TXT file format. Note that we set the image dimensions to 512x512 pixels and a maximum of 6000 batches.



(a) Map area selection



(b) Photo generation based on map selection

Fig. 7. Image grid generation tool.



Fig. 8. Example of labeling a training image.

3. MaskRCNN

For training with the MaskRCNN algorithm, a notebook within Google Colab [38] was used. The algorithm makes use of the Pascal VOC annotation format [39] in XML file format. The training parameters set were 150 epochs and minimum detection confidence of 70%. Note that the notebook used saves the trained model at each epoch, which allows us to choose the best model. The trained models that gave the best value in any of the metrics F1, Accuracy, Average Accuracy, True Positive, False Positive, False Negative are selected. Subsequently, the selected model has the best results in the evaluation process from among them.

4. Detectron2

Finally, for the training of the Detectron2 algorithm, a Google Colab workbook [40] has been utilized. It uses the COCO annotation format [41] in JSON file format. It is worth noting that, for this algorithm, we can include the segmentation annotations for each tag. In order to be on equal footing with the previous algorithms, we have not included segmentation annotations except the bounding box annotations.

This method only uses the algorithm Faster-RCNN, whose basic configuration file can be found in the Detectron2 Repository [42]. In addition, we set the learning rate to 0.01 with a maximum of 50,000 iterations and steps in 30,000, 40,000, and 45,000.

C. Legal Registration Check System

For the registration of properties, some organizations aim to guarantee the legal security of the operations carried out in the real estate market. For example, in Spain, the registration of a

swimming pool is not compulsory, but it is always advisable, as otherwise, it will not be valid in the eyes of third parties. In addition, the swimming pool affects the payment of real estate tax (IBI) since, like other types of constructions and installations on the property, it adds additional value to the property.

The name and correspondence of the bodies responsible for ensuring the registration of new works vary from country to country. For example, in the USA, the responsibility for these registrations belongs to the municipalities. In Spain, on the other hand, there is a body in charge of this competence called the Dirección General del Catastro (General Directorate of Cadastre). In both cases, these bodies offer a service to check the buildings registered for a given point on the coordinate map used in services such as Google Maps.

Therefore, we propose that the system obtains the coordinates corresponding to these pools after obtaining the pools detected in the images. In this way, the system will check if the property corresponding to the coordinates obtained has the detected pool registered. In this case, the system makes this check through a request to an endpoint of the API offered by the services of the ayliated organization. Based on the data obtained in response, it will be possible to check the registered constructions, determining whether or not the pool is registered.

VI. RESULTS

The tools used for training the models have methods for calculating some metrics using the validation set. However, we use a new process to evaluate the algorithms without relying on these tools and evaluating some new issues. This method consists of a new set, the evaluation set, which contains representative images of the final system. In addition to comparing the different models trained with the three algorithms, the method compares images with zoom 18 (Z18 set) and images with zoom 19 (Z19 set).

This Section is structured as follows: Section VI.A explains the metrics used. Section VI.B details the evaluation set. Sections VI.C, VI.D and VI.E show the results obtained from each algorithm. Section VI.F shows the comparison between the models trained with the different algorithms is given. Also, a comparison is also made between the Z18 set and the Z19 set. Finally, the VI.G section describes a system for check if the pools are registered legally by the appropriate agency.

A. Metrics

For the quantitative evaluation, we used the following metrics: accuracy, i.e., the relationship between true positives (TPs) and true positives (TPs) along with false positives (FPs) (Equation 1); recall, which is the probability that an image is classified as positive and

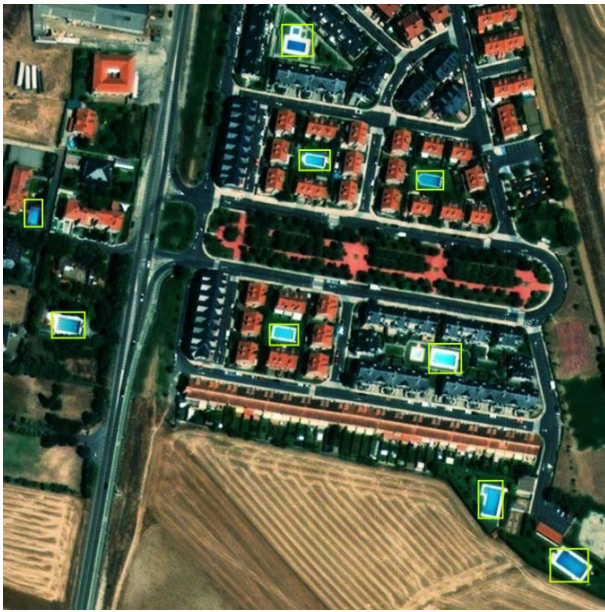


Fig. 9. Evaluation image with zoom 18.

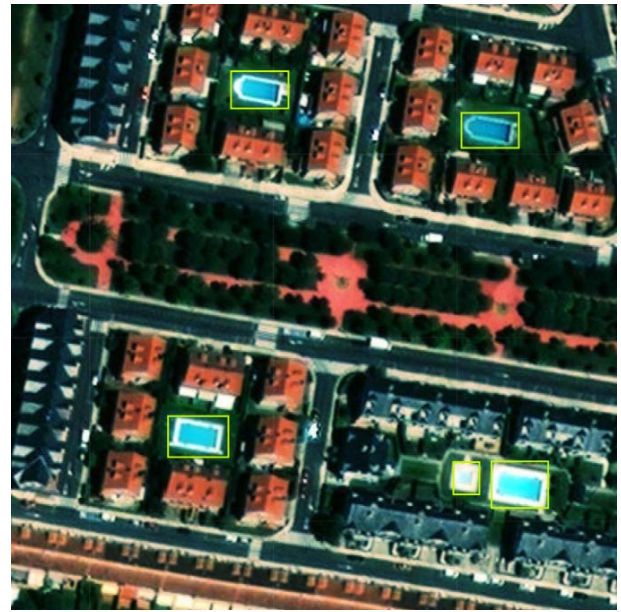


Fig. 10. Evaluation image with zoom 19.

the relationship between the TPs and the TPs together with the false negatives (FNs) (Equation 2); and F1, which is combination of the two previous metrics (Equation 3).

We classified the speed measured in frames per second (FPS); the mean average precision (mAP), calculated by the precision and recall curve; and the intersection over union (IoU), which is the overlapping area between the annotated bounding box of the object in the image and the detected bounding box by the model. For the mAP measure, the notation of mAP@X is used, where X indicates the IoU threshold value used to calculate the AP.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (3)$$

B. Evaluation Set

For the evaluation of the algorithms, we establish a reduced set of images. We have chosen eight locations in the province of Salamanca (Spain), obtaining a zoom 18 image (Z18 set) and a zoom 19 image (Z19 set) for each location. In total, we have obtained 16 images. Fig. 9 shows a Z18 image of a point and Fig. 10 shows a Z19 image, with their respective labelling.

We consider the following criterion for calculating the metric values for each trained model: It will consider neither a true positive nor a false positive if the model detects a not labeled pool for the evaluation.

These initially unlabelled objects are usually of tiny size in the image. Human experts are also not able to classify whether they are pools or not. This evaluation set contains in total 85 pools. Furthermore, the set splits into two image sets with 50 pools for the Z18 set and 35 pools for the Z19 set (Table III).

TABLE III. NUMBER OF POOLS IN EACH SET OF ASSESSMENT PICTURES

Set	Images	Pools
Zoom 18	8	50
Zoom 19	8	35
All	16	85

C. Yolov4

As can be seen in Fig. 11 the model trained with the YoloV4 neural network managed to detect 365 pools correctly and 56 detections as false positives using the validation set. This model has given 89.75% of mAP@0.50, 87% accuracy, 89% recall, and 88% F1-Score. These results come from a confidence threshold of 25%.

```

100
detections count = 820, unique_truth_count = 408
class_id = 0, name = Piscinas, ap = 89.76% (TP = 365, FP = 56)
for conf_thresh = 0.25, precision = 0.87, recall = 0.89, F1-score = 0.88
for conf_thresh = 0.25, TP = 365, FP = 56, FN = 43, average IoU = 61.39 %
IoU threshold = 50 %, used Area-Under-Curve for each unique Recall
mean average precision (mAP@0.50) = 0.897555, or 89.76 %
total Detection Time: 13 Seconds
    
```

Fig. 11. Validation results of the best model trained with YoloV4.

In Fig. 12 and Fig. 13 we can observe the detection on images of the evaluation set with zoom 18 and zoom 19, respectively. These example of detection use a confidence threshold of 10% as it has yielded the best results between the confidence thresholds 70%, 50%, 25%, and 10%.



Fig. 12. YoloV4 detection with zoom 18.



Fig. 13. YoloV4 detection with zoom 19.

The model trained with the zoomed image 18 has managed to detect 8 pools out of 9, one false positive, and there is one discarded detection as we do not know if it is a pool or not. As for the zoom image 19 it has detected 5 pools out of 5 and there is one discarded detection for the same reason.

Table IV shows the results obtained when using the evaluation set for detection with the model trained on YoloV4 with a confidence threshold of 10 percent. The nomenclatures used in the table are: **TP** = True Positive, **FP** = False Positive, **FN** = False Negative, **GT** = Ground Truth or total of truth pool, **Prec** = Precision, **RC** = Recall and **F1** = F1-Score.

TABLE IV. VALUES OF THE YOLOV4 MODEL METRICS ON THE EVALUATION SET

Set	TP	FP	FN	GT	Prec	RC	F1
Z18	43	2	7	50	95.6%	86.0%	90.5%
Z19	33	0	2	35	100%	94.3%	97.1%
All	76	2	9	85	97.4%	89.4%	93.3%

D. MaskRCNN

Training with MaskRCNN returned all models for each epoch. This algorithm allows us to evaluate each model resulting from each epoch individually. In this case, we have chosen the trained models whose results stand out in some of the metrics returned by the training tool with the validation set, such as mAP@0.5, precision, recall, and F1-Score. For each of these chosen models, we have calculated their metrics with the evaluation set. Among them, we highlight the model resulting from epoch 46 that gave us the best results. This trained model stood out from the other models resulting from this training because of its high accuracy, whose value reached 86.3%. In Fig. 14 we can observe the values of the metrics that the epoch 46 model has had with the validation set.

```

=====] - 90s 691ms/step - loss: 0.5755 - val_loss: 1.2065
ference model (last checkpoint of the train model)
map: 0.7354 TP: 297 FP: 47 FN: 111 total: 408 precision: 0.8633 recall: 0.7279 F1: 0.7898
    
```

Fig. 14. Validation results of the best model trained with MaskRCNN.

It has correctly detected 297 pools or true positives and incorrectly detected 47 pools or false positives. It has reached a mAP@0.5 of 73.54%, a recall of 72.79%, and an F1-Score of 78.98%.

In Fig. 15 and Fig. 16 we can observe the detections on the images of the evaluation set with zoom 18 and zoom 19, respectively.

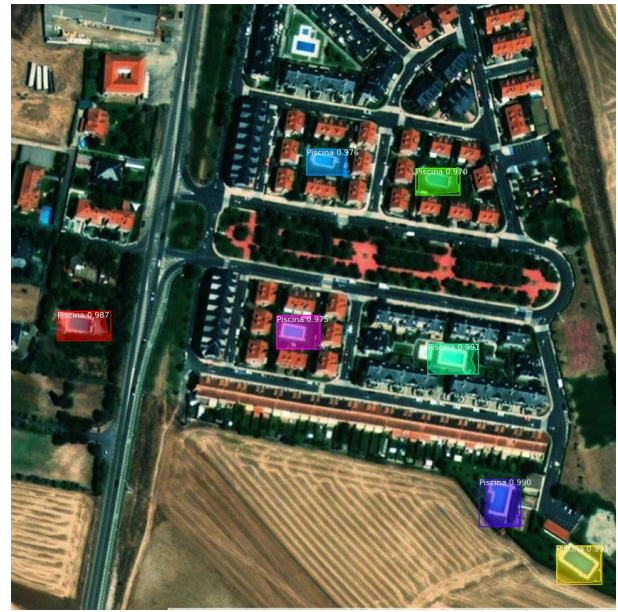


Fig. 15. Example of MaskRCNN detection with zoom 18.

These detections use a confidence threshold of 90% because it yields better results and allows filtering out more false positives. The model has detected 7 pools out of 9 in the zoom 18 image and 4 out of 5 in the zoom 19 image.



Fig. 16. Example of MaskRCNN detection with zoom 19.

Table V shows the results obtained when using the evaluation set for detection with the epoch 46 model trained on MaskRCNN with a confidence threshold of 90%.

TABLE V. VALUES OF THE METRICS OF THE MASKRCNN MODEL ON THE EVALUATION SET

Set	TP	FP	FN	GT	Prec	RC	F1
Z18	29	0	21	50	100%	58.0%	73.4%
Z19	29	2	6	35	93.5%	82.9%	87.9%
All	58	2	27	85	96.7%	68.2%	80.0%

E. Detectron2

Finally, the model trained with Detectron2 has obtained an AP@0.5 of 87.23% with the validation set, as we can see in Fig. 17. In this case, the tool used for training with Detectron2 also calculates the mAP@0.5:0.95 with a result of 35.06% and the AP@0.75 with 16.43%.

```

AP      AP50   AP75   APs     APm     AP1
-----|-----|-----|-----|-----|-----
35.056 | 87.238 | 16.431 | 20.691 | 36.411 | nan
[09/02 03:44:51 d2.evaluation.coco.evaluation]: Some metrics cannot be computed and is shown as NaN.
[09/02 03:44:51 d2.engine.defaults]: Evaluation results for my_dataset_val in csv format:
[09/02 03:44:51 d2.evaluation.testing]: copypaste: Task: bbox
[09/02 03:44:51 d2.evaluation.testing]: copypaste: AP, AP50, AP75, APs, APm, AP1
[09/02 03:44:51 d2.evaluation.testing]: copypaste: 35.0561,87.2379,16.4311,20.6906,36.4108,nan
    
```

Fig. 17. Validation results of the best model trained with Detectron2.

Fig. 18 and Fig. 19 show the detections on the images of the evaluation set with zoom 18 and zoom 19, respectively. These detections use a confidence threshold of 50%. The trained model detects 7 pools out of 9 in the zoom 18 image and 4 out of 5 in the zoom 19 image.

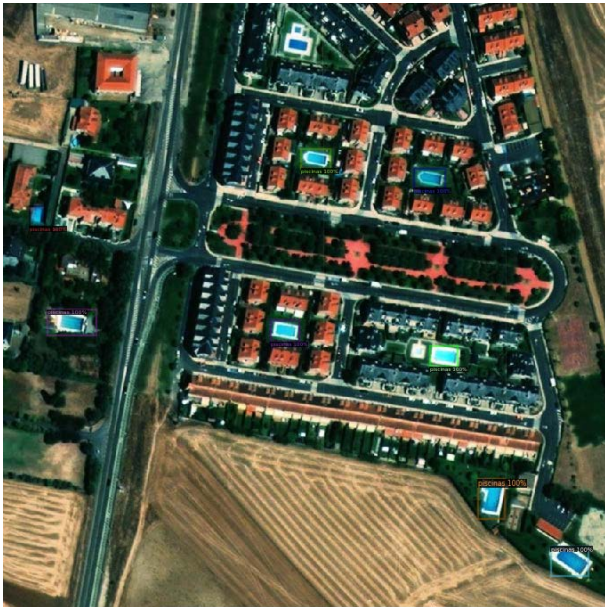


Fig. 18. Example of Detectron2 detection with zoom 18.

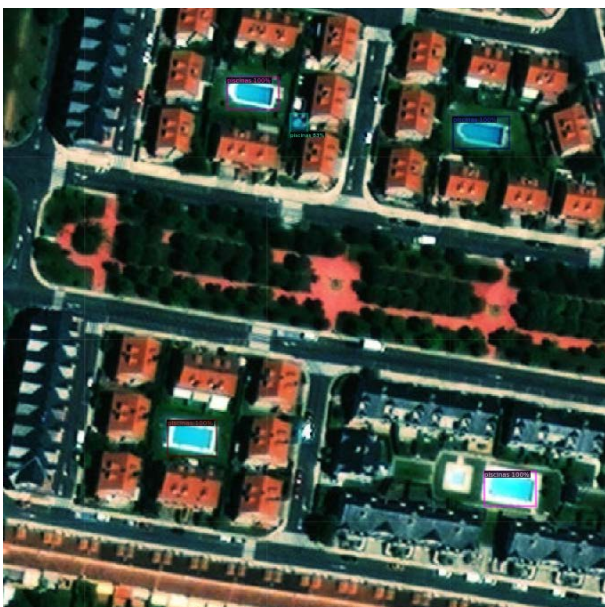


Fig. 19. Example of Detectron2 detection with zoom 19.

Table VI shows the results obtained when using the evaluation set for pool detection with the model trained with Detectron2 with a confidence threshold of 50%.

TABLE VI. VALUES OF THE METRICS OF THE DETECTRON2 MODEL ON THE EVALUATION SET

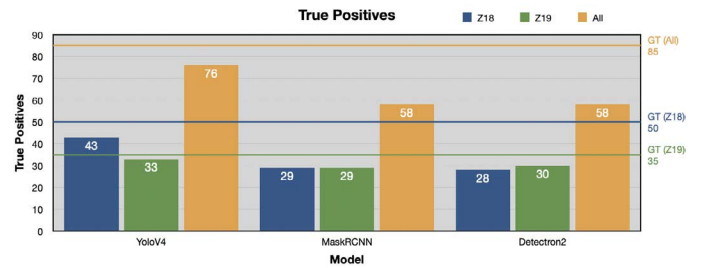
Set	TP	FP	FN	GT	Prec	RC	F1
Z18	28	5	22	50	84.8%	56.0%	67.5%
Z19	30	3	5	35	90.9%	85.7%	88.2%
All	58	8	27	85	87.9%	68.2%	76.8%

F. Comparison

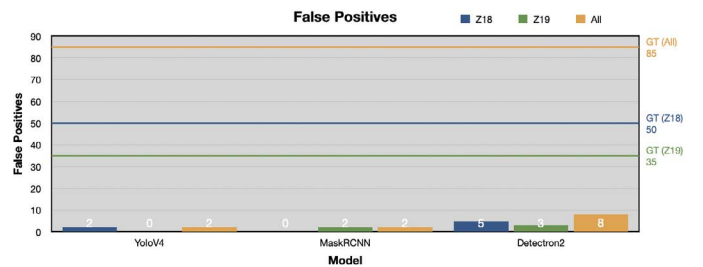
Once we have obtained the metrics of each model trained with the evaluation set, we proceed to compare the algorithms. Fig. 20a, Fig. 20b and Fig. 20c show plots comparing the models with the true positive, false positive and false negative values, respectively.

The model trained with the YoloV4 neural network yielded much better results than the other two algorithms. It managed to detect many more pools correctly, and, in addition, it only detected 2 pools incorrectly. On the other hand, if we use a confidence threshold of 25% or more with the model trained with YoloV4 we achieve that the model does not detect pools incorrectly. With the 25% threshold, we obtain 61 true positives and 0 false positives in the whole set.

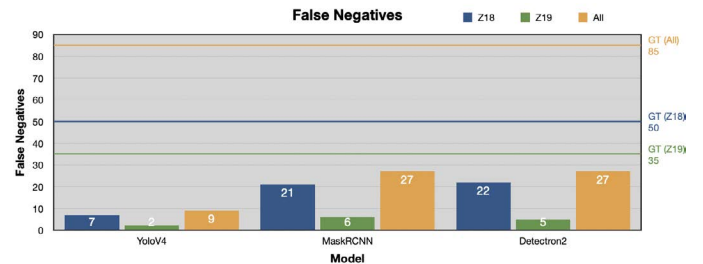
Finally, if we compare between the set of zoom 18 images (Z18) with the set of zoom 19 images (Z19) we can observe that in the three algorithms, there is a higher detection of true positives with the set Z19.



(a) True positives value comparison.



(b) False positive value comparison.



(c) False negative value comparison.

Fig. 20. True positive, false positive and false negative values comparison.

The YoloV4 algorithm has been able to detect 33 pools out of 35 in the Z19 set, while it has detected 43 out of 50 in the Z18 set. This difference is even more marked in the other two algorithms, where the number of true positives is almost identical between set Z18 and set Z19.

If we compare the precision metric of each model (Fig. 21) we obtain that the model trained with YoloV4 is more accurate than the other two algorithms. Furthermore, if we increase its confidence threshold to 25%, we obtain a precision of 100% on all three sets: Z18, Z19, and All. Finally, we can see that there is a lower precision with the Z18 set in all three algorithms compared to the Z19 set.

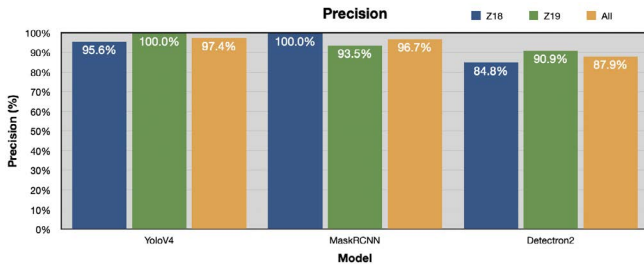


Fig. 21. Precision metrics comparison.

As for the recall metric, the YoloV4 algorithm has yielded better results, reaching 94.3% in the Z19 set. In addition, we see that with the Z19 set, they achieve a higher recall than the Z18 set, with a difference of almost 30% in the case of the Detectron2 algorithm.

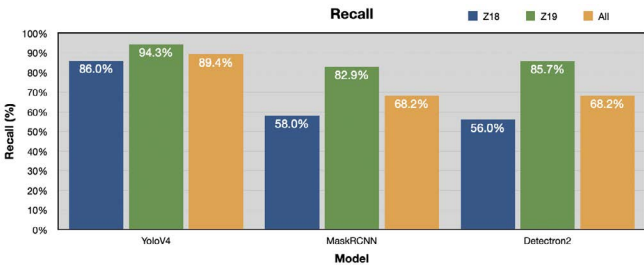


Fig. 22. Recall metric comparison.

Finally, regarding the F1-Score metric, which combines the precision and recall metrics, we can observe in Fig. 23. The YoloV4 algorithm is far superior to the other two algorithms, achieving up to 97.1% with the Z19 set. On the other hand, we see better results with the Z19 set compared to the Z18 set.

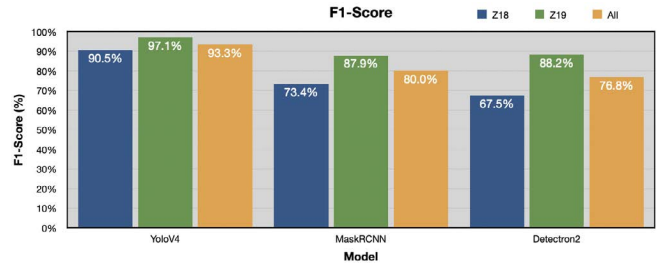


Fig. 23. F1-Score metric comparison.

G. Test Case of the Verification System

Before we can comment on the test case performed for the test system, we must describe it to know the system's situation. In this case, the test case has performed in a town near the city of Salamanca. The name of this town is Villamayor.

The *Dirección General del Catastro* is the body in charge of controlling, verifying and managing the records of buildings in Spain. The system has collected the images and check them, going through the pool detection subsystem. Their coordinates have been obtained from the detected pools to request the service offered by the *Dirección General del Catastro*. This legal system is in charge to check if the pools located in properties were registered. The result obtained from this process is in Fig. 24.

Fig. 24 shows three types of points distinguished by colors:

- **Blue:** This item identifies a pool detected by the system. This item has not yet gone through the system verification process.
- **Green:** This point identifies a pool detected in the system and was verified as registered to the corresponding property through its geographical coordinates.



Fig. 24. Results from the verification system.

- **Red:** This item identifies a swimming pool detected in the system and has no record for the corresponding property with its geographical coordinates.

In the observed village of Villamayor, as observed in Fig. 24, there are actually 27 pools. Of those 27 pools, the system has been able to detect 23 pools and unable to detect 5 pools that actually exist. Of those 23 detected pools, 22 of them are correct, however, one of them is incorrect. Of these pools, 22 are registered and linked to the related property, one is not registered, and one is still pending verification.

VII. CONCLUSIONS

This paper demonstrates that it is possible to determine the presence of a pool in an image with an accuracy better than 97% using a multi-agent architecture that allows distributed computing and has allowed the evaluation of different algorithms combined to improve the detection process.

After evaluating the algorithms and comparing them, we highlight the model trained with the YoloV4 neural network, which offers better results in all metrics. We propose to use this algorithm for pool detection using a confidence threshold of 10% in case errors or false positives can be assimilated, allowing a higher detection or recall, or to use a confidence threshold of 25% in case higher accuracy in pool detection is sought. Finally, it has been observed that for the detection of pools, it is better to use images with zoom 19 versus zoom 18. Although for images with zoom 19, it is necessary to process four times more images for the same area than with images with zoom 18, it is very convenient to sacrifice more computational resources to detect pools. Moreover, in this case, study, it is not vital to display the detections in real-time, so spending a few extra seconds in the detection process is not a concern.

Finally, a test case is made to observe a specific population to check the system's operation. In this way, based on the test carried out in the town of Villamayor, an operating system has been verified. This test has been possible to certify the system's effectiveness to determine which pools are registered or not in the corresponding official bodies. In addition, the system has proven to help check the automatic detection of pools and the checking of pool records.

ACKNOWLEDGMENTS

Héctor Sánchez San Blas's research is supported by the Spanish Ministry of Universities (FPU Fellowship under Grant FPU20/03014). The research of Luis Augusto Silva has been funded by the call for predoctoral contracts USAL 2021, co-financed by Banco Santander.

REFERENCES

- [1] D. Tien, T. Rudra, A. B. Hope, "Swimming pool identification from digital sensor imagery using SVM," *Proceedings - Digital Image Computing Techniques and Applications: 9th Biennial Conference of the Australian Pattern Recognition Society, DICTA 2007*, pp. 523–527, 2007, doi: 10.1109/DICTA.2007.4426841.
- [2] W. Passos, E. Silva, S. Netto, J. Martins, Y. Costa, G. Araujo, A. Lima, "Detecção de Potenciais Focos do Aedes aegypti em Vídeos Aéreos Usando Redes Neurais," pp. 22–25, 2020, doi: 10.14209/sbrt.2020.1570661555.
- [3] P. C. Gray, K. C. Bierlich, S. A. Mantell, A. S. Friedlaender, J. A. Goldbogen, D. W. Johnston, "Drones and convolutional neural networks facilitate automated and accurate cetacean species identification and photogrammetry," *Methods in Ecology and Evolution*, vol. 10, no. 9, pp. 1490–1500, 2019, doi: 10.1111/2041-210X.13246.
- [4] M. I. Habibie, T. Ahamed, R. Noguchi, S. Matsushita, "Deep Learning Algorithms to determine Drought prone Areas Using Remote Sensing and GIS," pp. 69–73, 2020, doi: 10.1109/AGERS51788.2020.9452752.
- [5] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [6] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 580–587, 2014, doi: 10.1109/CVPR.2014.81.
- [7] R. Girshick, "Fast R-CNN," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, pp. 1440–1448, 2015, doi: 10.1109/ICCV.2015.169.
- [8] H. Rampersad, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Total Performance Scorecard*, pp. 159–183, 2020, doi: 10.4324/9780080519340-12.
- [9] K. He, G. Gkioxari, P. Dollár, R. Girshick, "Mask r-cnn," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2020, doi: 10.1109/TPAMI.2018.2844175.
- [10] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, K. He, "Detectron." <https://github.com/facebookresearch/detectron>, 2018.
- [11] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, R. Girshick, "Detectron2." <https://github.com/facebookresearch/detectron2>, 2019.
- [12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, A. C. Berg, "SSD: Single Shot MultiBox Detector," 12 2015, doi: 10.1007/978-3-319-46448-0_2.
- [13] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You only look once: Unified, real-time object detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 779–788, 2016, doi: 10.1109/CVPR.2016.91.
- [14] J. Redmon, A. Farhadi, "YOLO9000: Better, faster, stronger," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 6517–6525, 2017, doi: 10.1109/CVPR.2017.690.
- [15] J. Redmon, A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018, doi: <http://arxiv.org/abs/1804.02767>.
- [16] A. Bochkovskiy, C. Wang, H. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *CoRR*, vol. abs/2004.10934, 2020, doi: <https://arxiv.org/abs/2004.10934>.
- [17] S.-H. Chen, C.-W. Wang, I.-H. Tai, K.-P. Weng, Y.-Chen, K.-S. Hsieh, "Modified yolov4-densenet algorithm for detection of ventricular septal defects in ultrasound images," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 7, pp. 101–108, doi: 10.9781/ijimai.2021.06.001.
- [18] L. A. Silva, H. S. S. Blas, D. P. García, A. S. Mendes, G. V. González, "An architectural multi-agent system for a pavement monitoring system with pothole recognition in uav images," *Sensors*, vol. 20, no. 21, pp. 1–23, 2020, doi: 10.3390/s20216205.
- [19] C. Galindo, P. Moreno, J. Gonzalez, V. Arévalo, "Swimming pools localization in colour high-resolution satellite images," vol. 4, pp. IV–510–IV–513, 2009, doi: 10.1109/IGARSS.2009.5417425.
- [20] M. Kim, J. B. Holt, R. J. Eisen, K. Padgett, W. K. Reisen, J. B. Croft, "Detection of swimming pools by geographic object-based image analysis to support west Nile virus control efforts," *Photogrammetric Engineering and Remote Sensing*, vol. 77, no. 11, pp. 103–113, 2011, doi: 10.14358/pers.77.11.1169.
- [21] B. Rodríguez-Cuenca, M. C. Alonso, "Semi-automatic detection of swimming pools from aerial high-resolution images and LIDAR data," *Remote Sensing*, vol. 6, no. 4, pp. 2628–2646, 2014, doi: 10.3390/rs6042628.
- [22] C. Ferner, G. Eibl, A. Unterweger, S. Burkhart, S. Wegenkittl, "Pool detection from smart metering data with convolutional neural networks," *Energy Informatics*, vol. 2, pp. 1–9, 2019, doi: 10.1186/s42162-019-0097-8.
- [23] Z. Domozi, A. Molnar, "Surveying private pools in suburban areas with neural network based on drone photos," *EUROCON 2019 - 18th International Conference on Smart Technologies*, pp. 1–6, 2019, doi: 10.1109/EUROCON.2019.8861770.
- [24] B. Lima, L. Ferreira, J. M. Moura, "Helping to detect legal swimming pools with deep learning and data visualization," *Procedia Computer Science*, vol. 181, no. 2019, pp. 1058–1065, 2021, doi: 10.1016/j.procs.2021.01.301.
- [25] C. Zato, G. Villarrubia, A. Sanchez, I. Barri, E. Rubión, A. Fernández, C. Sánchez, J. Cabo, T. Álamos, J. Sanz, J. Seco, J. Bajo, J. Corchado Rodríguez, "Pangea - platform for automatic construction of organizations of intelligent agents," vol. 151, 01 2012, doi: 10.1007/978-3-642-28765-7_27.
- [26] J. Schwartz, et al., "Bing maps tile system," 2009. <http://msdn.microsoft.com/en-us/library/bb259689.aspx>, (accessed in: 13/09/2021).
- [27] "Google maps." <https://maps.google.com>, (accessed in: 13/09/2021).

- [28] “Open street maps.” <https://www.openstreetmap.org/>, (accessed in: 13/09/2021).
- [29] “Esri world imagery.” <https://www.arcgis.com/apps/mapviewer/index.html?layers=10df2279f9684e4a9f6a7f08febac2a9>, (accessed in: 13/09/2021).
- [30] “Wikimedia maps.” <https://maps.wikimedia.org/>, (accessed in: 13/09/2021).
- [31] “Nasa gibs.” <https://map1.vis.earthdata.nasa.gov/>, (accessed in: 13/09/2021).
- [32] “Carto light.” <https://cartodb-basemaps-c.global.ssl.fastly.net/>, (accessed in: 13/09/2021).
- [33] “Stamen toner b & w.” <https://stamen.com/>, (accessed in: 13/09/2021).
- [34] “Sentinel.” <https://www.sentinel-hub.com/>, (accessed in: 13/09/2021).
- [35] “Roboflow.” <https://roboflow.com>, (accessed in: 13/09/2021).
- [36] “Swimming Pool Detect.” <https://github.com/Hectorssb/SwimmingPoolDetection>, (accessed in: 02/11/2022).
- [37] “YoloV4 cfg.” <https://github.com/Hectorssb/SwimmingPoolDetection/blob/main/Yolov4/cfg/yolov4-obj.cfg>, (accessed in: 02/11/2022).
- [38] “Train Mask-RCNN Model on Custom Data.” <https://colab.research.google.com/drive/1rBuhT8AjP2td20otdUnpuF7CCMmjRb4O>, (accessed in: 13/09/2021).
- [39] M. Everingham, L. Van Gool, C. K. I. Williams, Winn, A. Zisserman, “The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.” <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>.
- [40] “Detectron2 Beginner’s Tutorial.” https://colab.research.google.com/drive/1n_nulKMxxCF6Jg4WMw20mO8R6rqC078, (accessed in: 13/09/2021).
- [41] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, P. Dollár, “Microsoft coco: Common objects in context,” 2014. [Online]. Available: <https://arxiv.org/abs/1405.0312>, doi: 10.48550/ARXIV.1405.0312.
- [42] Detectron2, “faster_rcnn_X_101_32x8d_FPN_3x.” https://github.com/facebookresearch/detectron2/blob/main/configs/COCO-Detection/faster_rcnn_X_101_32x8d_FPN_3x.yaml, (accessed in: 13/09/2021).



Héctor Sánchez San Blas

He is a Researcher at the University of Salamanca. He is the beneficiary of an FPU predoctoral contract in the 2020-2021 call. He is a Ph.D. student in computer engineering at the same university and has a degree in Computer Engineering and a Master’s degree in Intelligent Systems at the same center. During his master’s degree, he was a collaborating researcher at the Expert Systems and Applications Laboratory (ESALab) of this university, collaborating with research projects related to the Internet of Things, Virtual Reality applications, and machine learning. Currently, he is researching the development of IoT and neural networks together with research projects focused on machine vision and Smart Cities.



Antía Carmona Balea

She is a Pd.D. student in Computer Engineering at University of Salamanca (USAL). She has a degree in Chemistry at the University of Valladolid (UVA), she received the master’s degree in Meteorology from UNED and she is currently doing a master in scientific communication. Her doctoral studies deal with IoT and artificial intelligence which she combines with her work as a secondary school teacher.



André Sales Mendes

He received a master’s degree in Intelligent Systems and the Ph.D. degree in Computer Engineering from University Of Salamanca (USAL). He is current working on Expert System And Applications Laboratory Research Group, Computer Science Department, University of Salamanca, as an Adjunct Professor. It has a numerous publication in international impact journals indexed in JCR reference ranking. His main research interests focus on the field of artificial intelligence, IoT (internet of things) and robotic. He also collaborates in several research projects within the research group.



Luís Augusto Silva

He received his Master’s degree in Applied Computing from the University of Itajaí Valley, Brazil, in 2019. He has a degree in Internet Systems from the Federal Institute of Santa Catarina (IFC), Camboriú, Brazil, ending in February 2017. His research during his master’s degree covered the field of Notification Systems, IoT, and Data Privacy. During the master’s degree, he was a collaborating researcher at the Laboratory of Embedded and Distributed Systems (LEDS) at UNIVALI, collaborating with research projects related to the Internet of Things. Since August 2020, he has been a Ph.D. student in Computer Engineering at the Universidad de Salamanca - Spain, and a researcher at the Expert Systems and Applications Laboratory (ESALab). His research lines are directly related to Internet of Things, Embedded Drone Systems, and Data Privacy applied to Smart Environments.



Gabriel Villarrubia González

He received the master’s degree in intelligent systems from the University of Salamanca, in 2012, the master’s degree in Internet Security, in 2014, the master’s degree in information systems management, in 2015, and the Ph.D. degree from the Department of Computer Science and Automation, University of Salamanca. He is a Computer Engineer at the Pontifical University of Salamanca in 2011. He is currently a Research Professor with the Department of Informatics. Throughout his training, he has followed a well-defined line of research, focused on the application of multi-agent systems to ambient intelligence environments, with special attention to the definition of intelligent architectures and the fusion of information.

Emotion-Aware Monitoring of Users' Reaction With a Multi-Perspective Analysis of Long- and Short-Term Topics on Twitter

Danilo Cavaliere¹, Giuseppe Fenza^{1*}, Vincenzo Loia¹, Francesco Nota²

¹ Dipartimento di Scienze Aziendali e Management Information Systems (DISA-MIS), University of Salerno, Salerno (Italy)

² Defence Analysis & Research Institute, Center for Higher Defence Studies, 00165 Rome (Italy)



Received 28 October 2022 | Accepted 23 December 2022 | Published 1 February 2023

ABSTRACT

Social networks, such as Twitter, play like a disinformation spread booster giving the chance to individuals and organizations to influence users' beliefs on purpose through tweets causing destabilization effects to the community. As a consequence, there is a need for solutions to analyse users' reactions to topics debated in the community. To this purpose, state-of-the-art methods focus on selecting the most debated topics over time, ignoring less-frequent-discussed topics. In this paper, a framework for users' reaction and topic analysis is introduced. First the method extracts topics as frequent itemsets of named entities from tweets collected, hence the support over time and RoBERTa-based sentiment analysis are applied to assess the current topic spread and the emotional impact, then a time-grid-based approach allows a granule-level analysis of the collected features that can be exploited for predicting future users' reactions towards topics. Finally, a three-perspective score function is introduced to build comparative ranked lists of the most relevant topics according to topic sentiment, importance and spread. Experiences demonstrate the potential of the framework on IEEE COVID-19 Tweets Dataset.

KEYWORDS

Frequent Itemsets, Multi-perspective Topic Monitoring, Sentiment Analysis, Users' Reaction Prediction.

DOI: 10.9781/ijimai.2023.02.003

I. INTRODUCTION

In these days, fast and easy access to the Internet allows everyone to express their own ideas about different topics (politics, social events, etc.) by writing posts on social networks (e.g., FB posts, tweets) or even through their own blog or site that can reach millions of people worldwide. On one hand, this phenomenon certainly encourages the freedom of speech, but on the other hand, it leaves society vulnerable and helpless against possible news manipulation coming from multiple sources. Recent examples are the large-scale disinformation and misinformation related to Brexit and US presidential campaign in 2016 or the global infodemic associated to the Covid-19 pandemic and vaccine campaign. Social networks increase the effectiveness and scale of disinformation, that is tailor-made to manipulate users' beliefs on purpose by exploiting persuasive and propaganda techniques. Real and fake news attract users' attention by leveraging their psychological and emotional states leading them to react by expressing their own opinions on social networks. Among the most popular social networks, Twitter is vastly adopted to express reactions towards news and events, therefore, tweet mining could serve as a tool to assess the public opinion about news. Countering information

disorder requires monitoring the evolution of trendy topics in order to early alert policymakers and governments about potential risks enabling them to mitigate the impact and harms. Therefore, automatic tools for preventing the spread of news are demanded to effectively contrast the phenomenon that can lead to strong disagreement among people and violent protests as an effect in the worst cases.

Canonical disinformation (e.g., fake news and hoaxes), as well as propagandistic tweets, appeal to user's sentiments to influence his/her opinions with the aim of creating confusion or satisfying the writer's intentions that may be of a different kind (e.g., political, social, economic, etc.). As a consequence, the analysis of people's emotional reactions can help monitoring strong reactions towards certain news and act as an early-stage signal to prevent massive disinformation spread. Since news is meant to provoke emotional reactions, the analysis of the emotional aspect is crucial. For this purpose, Sentiment Analysis provides tools to identify, extract, quantify, and study affective states and subjective information by using computational and linguistics techniques, including natural language processing, text analysis, computational linguistics, and biometrics [1].

Beyond the single-post linguistic analysis, time-sensitive, continuous, and heterogeneous information spreading should be constantly analysed since it represents one of the primary issues to the development of good-performing online information monitoring systems [2]. As a consequence, to fight against online fake news spreading, models and new monitoring tools are required to capture

* Corresponding author.

E-mail address: gfenza@unisa.it (Giuseppe Fenza).

the dynamic nature of online information to promptly detect eventual attacks aimed at generating cognitive vulnerabilities and society destabilization effects.

For this purpose, a crucial point to online information monitoring is the analysis of topics debated over time in a community, by taking into consideration not only those ones discussed for long time, but also those topics debated in a restricted period of time, that may influence users' behaviours and thinking as well as the long-time-debated ones. This paper presents an emotion-aware solution to analyse users' reactions towards topics that have been constantly discussed over time (long-term topics) and topics that have been discussed in a specific brief period (short-term topics). The rationale behind the approach is to combine the emotional analysis of tweet content with the time frequent analysis of relevant topic itemsets and tweet spread to better evidence those topics that may have the strongest impact on the community. In other words, a mechanism is introduced to rank topics that may cause community destabilization effects by jointly considering topic sentiment, importance and spread. In detail, the approach collects tweets day-by-day in a reference period, extracts short- and long-term itemsets of topics from tweets, then evaluates topic mentions and extracts a sentiment score on four different emotional classes to depict users' reactions to topics. The approach is based on a granular time-grid-based data processing schema, that allows the emotion-aware analysis of the extracted short- and long-term topics which can be used for community monitoring, including the prediction of emotional users' reactions towards topics. The final and ultimate goal of the framework is a multi-perspective topic relevance analysis to provide ranked lists of topics in accordance with topic sentiment, importance, and spread.

In a nutshell, the paper contribution can be summarized in the following points:

- A report on multi-class emotional analysis of Twitter users' reactions showing that short-living topics, which are often discarded, may generally cause great emotional effects on community.
- The proposal of a time-grid-based approach to track topic mentions and their emotional impact over time, aimed at helping the detection of high-impact topics.
- The design of a time granular emotion-aware topic modeling to serve the collected information reuse for different tasks, including the prediction of eventual future users' reactions.
- The introduction of a score function combining topic sentiment, mention frequency and spread to perform a multi-perspective topic relevance analysis by comparing score-based topic impact ranked lists.

The paper is organized as follows: Section II provides the preliminaries to the research and discusses the related work, Section III describes the approach in detail, Section IV shows a case study related to the proposed approach, and Section V reports on the test conducted and conclusions close the paper.

II. PRELIMINARIES

A. A preliminary Study

As a preliminary step, a careful analysis of Twitter users' emotional reactions has been carried out with the aim of finding out meaningful features of the topics debated, the intensity of the emotions they have aroused in the community and the durability of these emotional reactions. The rationale behind this preliminary study is to determine the effective impact that some short-term topics may have on the community and whether they are of interest for analysis of users' future behaviors and mitigation action planning.

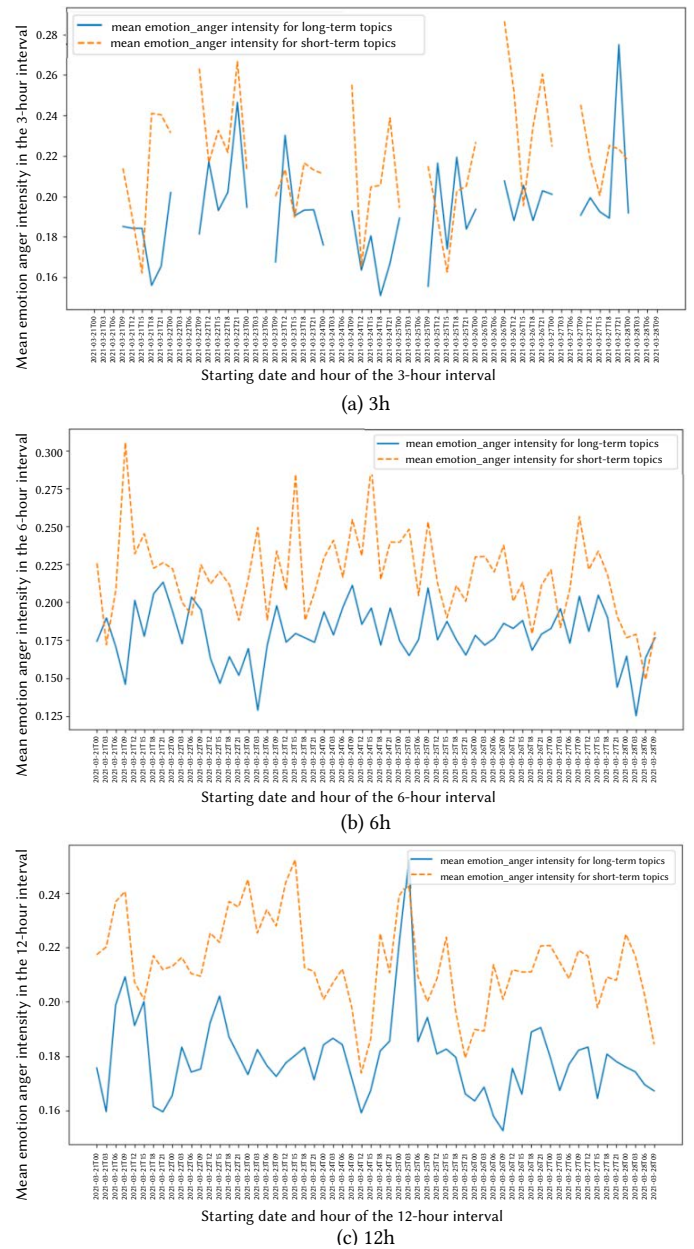


Fig. 1. Twitter users' averaged sentiment scores on anger class at different hour.

To analyse Twitter users' emotional reactions, tweets have been collected on a 3-months period from February 2021 to April 2021, hence tweet content has been extracted and processed with Sentiment Analysis to determine the emotional intensity (or score) on four emotions: sadness, anger, joy, and optimism. The sentiment scores have been analysed over time by using different time units, in detail the average sentiment score for each emotional class has been assessed each 3, 6, and 12 hours on short- and long-term topics, distinctively. This way, a time granular analysis of the emotional reactions has been conducted to evidence the emotional impact related to the debates on short-term topics.

Since from the analysis of the four emotional classes we can state analogous conclusions, for sake of simplicity results are reported for the emotional class anger in Fig. 1 on a period going from 03/04/2021 to 20/04/2021. The figure shows the average sentiment scores of the tweets associated with topics for the anger class on the short- and long-term topics as the 3-hour, 6-hour, and 12-hour analyses. As a general trend, let us say that short-term topics cause higher peaks in negative emotions (i.e., high curve peaks). In some emotional

classes, the sentiment score of some emotions on short-term topics is constantly higher than the long-term topic sentiment score. For example, long-term topic mean anger intensity is statistically lower than short-term topics mean anger intensity (with a p-value less than 0.00001). This result proves that the undoubted influence of short-term topics on users' emotional reactions may lead the community to suffer from emotional destabilization phenomena and that monitoring emotions is fundamental to plan mitigation tools for economic and political analyses.

From the figure, let us notice that the 6-hour interval analysis uniformity seems to be a rational trade-off between the 3-hour analysis (excessive variability) and the 12-hour analysis (accentuated flatness).

B. Related Work

The spread of misinformation is generally related to hot topics (e.g., Covid-19 virus spread and vaccines), that are subject of different analyses aimed at finding out important trends, such as a decline in the number of vaccine supporters caused by the spread of fake news on vaccines [3]. Many works in the literature focus on misinformation and disinformation detection from texts [4]–[7]. In [4], the authors present a method to label a dataset on propagandistic text, run topic modelling, and then corpora imbalance assessment for propaganda detection. In [5], different techniques (e.g., GloVe, BERT, and LSTM) are combined to perform word representation, pretrain the model and detect persuasive text. Another approach [6] allows fragment-level text analysis which exploits tf/idf, word, and character n-grams to build a classifier for propagandistic text. In [7], the authors present a Machine Learning framework for article- and sentence-level persuasive text detection. Other works analyse disinformation on social networks to help find countermeasures, such in [8] where the authors propose a framework exploiting activity-connectivity maps based on network and temporal activity patterns to detect social influence among ISIS supporters.

Another predominant vein in literature is aimed at finding solutions to deal with fake news detection from social networks. In [9], the authors collect news from heterogeneous sources and test out different Machine Learning methods for fake news detection. Some works investigate sentiment analysis, such as the method proposed in [10], which explores sentiment analysis and determines the most relevant elements for fake news detection, including multilingualism, explainability, bias mitigation, and multimedia element treatment. In [11], the authors proposed an attention-based approach for multi-modal sentiment analysis. In [12], sentiment extracted from news is coupled with domain parameters to improve predictions. Several works focus on Twitter misinformation spread monitoring, such as the approach proposed in [13], which checks emotional valence in relation to false stories certified by Google Fact Checker API on Covid-19, and finds out that the emotional valence varies depending on the different topics. Another work [14] analyses changes in Twitter users' behaviours after a misinformation attack by analysing variations with respect to frequency and sentiment expressed in their tweets. The solution proposed in [15] performs topic identification, network analysis, and sentiment analysis to classify tweets into the six categories for misinformation analysis. It comes out that sentiment score is mainly influenced by government measures and public speeches from government officials, as well as news agencies and public figures. In detecting misinformation, Sentiment Analysis is applied to news and social network posts with different intents, such as the approach presented in [16], which assesses text sentiment-harmful news correlation and detects harmful news through sentiment analysis.

Some other works are focused on Twitter data analysis aimed at extracting meaningful patterns for users' behavior extraction or prediction, such as the work proposed in [17], which presents a Latent Dirichlet Allocation-based model to analyse people's reactions to

Covid-19 from tweets, the method proposed in [18] that introduces a method to mine association rules from tweets and extract people's attitudes to topics, and the approach proposed in [19] that coupled topic identification and sentiment analysis to extract emotional people reactions from multi-lingual tweets. All these methods analyse exclusively general macro topics (e.g., politics, economic impacts, etc.) and focus on time evaluation that missed the momentary strong emotional impact that a topic may have on communities, and consequently fail in capturing a reliable topic relevance evaluation. To deal with this challenge, a granular time-grid hierarchical approach has been designed to process tweets, extract long-time and short-time-debated topics alongside their parameters (sentiment, topic mention frequency, tweet count), and evaluate them in time granules (grid cells) in order to achieve a better topic relevance assessment.

III. THE METHODOLOGY

The framework has been designed as a three-tier model allowing sentiment-aware topic extraction, time granular emotion-aware analysis of users' reaction to topics and, a three-perspective topic relevance analysis. The complete pipeline is shown in Fig. 2, where the first tier includes tasks to perform *Tweet collection and topic extraction*, which lead to building datasets of tweets collected, pre-processing their content to extract relevant entities by running Named Entity Recognition (NER), and extract topics as frequent word itemsets from tweets assessing their mention frequency over time. Then, the second tier allows *Users' reaction modeling* by accomplishing several tasks, including topic sentiment analysis aimed at depicting users' reaction towards itemsets and a time-granular emotion modeling that can serve the prediction of users' future reactions towards topics.

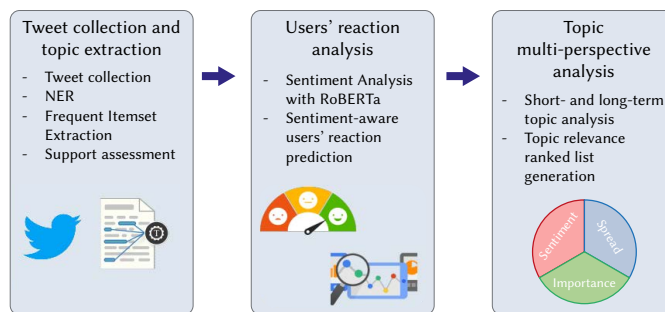


Fig. 2. Architecture pipeline.

The third tier is in charge of *Topic multi-perspective analysis* by running several tasks aimed at analysing the relevance of short- and long-term topics in the community by considering topic sentiment, importance, and spread in the community.

A. Tweet Collection and Topic Extraction

Tweets are collected to build a tweet dataset representing a multiplicity of topics. To polish the data acquired, widely-used Python libraries are exploited to perform state-of-the-art pre-processing tasks, including stop word, special characters and punctuation removal, tokenization, and Part-of-Speech (POS) tagging by using nltk¹, spaCy² and stanza³. The preprocessing task is closed by removing URLs from texts. To detect the effective entities debated in tweets, Named Entity Recognition (NER) is performed. Therefore, Stanza NER is then applied to hashtags and keywords to extract relevant names of people, organizations and locations discussed in tweets.

¹ <https://www.nltk.org/>

² <https://spacy.io/>

³ <https://stanfordnlp.github.io/stanza/>

Then, topics are extracted and discriminated into two specific categories: short- and long-term topics; the former refers to those topics that are constantly debated over time (high mention frequency over time), and the latter represents those topics that have been debated within a limited time interval (i.e., they generally last few hours, or a day at maximum, or they are debated only a few times in different time periods). Since short-term topics seem to cause the strongest emotional reactions among Twitter users, as has been found out in the preliminary study (Section A), the main rationale behind short-term/long-term discrimination is to detect those events causing the strongest users' reactions that, as a consequence, may lead to community destabilization effects.

In order to define topics, frequent word itemsets are built as composed of hashtags, keywords, mentions, and NEs extracted from tweets. In detail, itemsets are defined as 2-grams and 3-grams representing debated topics in the community by using the Python library `fpgrowth`. Then the support, defined as the occurrence frequency of a pattern in a dataset, is calculated. In this work the patterns are 2-grams and 3-grams, therefore the support is their occurrence frequency in the tweets dataset [20].

To allow a more robust topic analysis over time, the features of the extracted itemsets (e.g., support, tweet frequency, etc.) have been analysed by means of time granules. For this purpose, a time grid schema is defined to analyse the itemset support with respect to time at this stage, therefore, the time interval considered for tweet collection is represented as a grid having cells of fixed size expressed in hours. For instance, a 6-hours-cell in the time grid will report the time support for each itemset decomposed in each of the 6 hours of the time interval considered. The grid helps focus on the most debated topics in a specific moment in time (i.e., itemsets with high support in specific grid cells). To discriminate between long- and short-term itemsets/topics, a support-based filtering function is applied to itemsets separating those itemsets debated over a long time from those that have a localized time occurrence. The support-based filtering function is reported as Algorithm 1; it takes the topic T , a counter C and the time grid cell length in hours (H), hence it discards T if it has a low support (lines 3-5) otherwise it increases C by 1 or decreases it by $H/12$ depending on T support in each cell (lines 6-12). Then, it checks whether T is a short- or a long-term topic by applying the sigmoid function to the base two logarithm of counter C (lines 13-17). The rationale behind the logarithmic function is that it grows slowly allowing topics to be considered as long-term only if they are present (i.e. meaning their support is high) in many time-grid cells. Moreover, when a long-term topic becomes less present in tweets, the logarithm ensures a slow and gradual decrease when applied on the counter. Therefore, the longer the former long-term topic was discussed, the more time it will take to become short-term due to the decrease strategy. Conversely, if a topic has been debated for a while, but not for so long (e.g. one month), when it is not debated anymore it will come back faster to the state of short-term topic (or absent state in case it is not referred to at all). The sigmoid function helps achieve a clearer interpretation of the logarithmic function applied to the counter by forcing the score to lie within the range 0.5 to 1. For this reason, a threshold fixed to the half of the range (i.e. 0.75), has been used to discriminate between short- and long-term topics.

B. Users' Reaction Prediction

The second tier allows topic sentiment analysis and the prediction of users' emotional reactions towards the extracted topics.

For this purpose, first, Sentiment Analysis is exclusively applied to tweets related to topics selected through Algorithm 1. The sentiment analysis outcome on these tweets will be used to perform the sentiment analysis of the selected itemsets. To perform Sentiment Analysis, our

Algorithm 1. Support-based filtering on itemsets

```

1: Let  $T$  be a topic,  $C$  a counter and  $H$  the number of hours of the
   interval
Require:  $T \geq 0$ ,  $C \geq 0$  and  $H \geq 3$ 
Ensure:  $T$  is a short-term or long-term topic
2: function SUPPORT_FILTER( $T, C$ )
3:   if SUPPORT( $T$ )  $\leq 0.001$  then
4:      $T$  is discarded
5:   end if
6:   for each time grid cell do
7:     if SUPPORT( $T$ )  $> 0.005$  then
8:        $C \leftarrow C + 1$ 
9:     else if SUPPORT( $T$ )  $\leq 0.005$  then
10:       $C \leftarrow C - \frac{H}{12}$ 
11:    end if
12:  end for
13:  if SIGMOID( $\log_2 C$ )  $\leq 0.75$  then
14:     $T$  is a short-term topic
15:  else if SIGMOID( $\log_2 C$ )  $\geq 0.75$  then
16:     $T$  is a long-term topic
17:  end if
18: end function

```

framework employs RoBERTa [21], the robustly optimized pretraining approach of the famous NLP transformer-based machine learning technique Google BERT. RoBERTa-based model allows text Sentiment Analysis in terms of polarity classification and single emotional class classification, namely optimism, joy, sadness and anger. Once, RoBERTa Sentiment Analysis has been applied to tweets in a specific time-grid cell of duration H (e.g. 6 hours), the sentiment score for a topic T is calculated as the mean of the sentiment scores achieved for each tweet in which T is present. This is done for each of the four RoBERTa emotional classes, hence for each of the four emotions describing users' emotional reactions and for each time-grid an intensity score is associated for topic T . For example, the intensity scores are 160 if there are 4 time grid cells, which length is 6-hours each, in a total period of 10 days considered for all the 4 emotions (i.e. $4 \times 10 \times 4 = 160$).

To deal with the analysis of eventual users' reactions towards topics, a regressor module has been designed to process sentiment related to itemsets and accordingly predict how community users react to them. The regressor has been designed on the time grid introduced before (see Section A), where the support and sentiment on four classes are reported for each itemset in consecutive s -hour time-grid cells, where s is the number of hours considered as a step. The mean value of each emotional class and support is computed in each step. Then, given the mean sentiment value for each emotional class in the current and previous steps, the regressor goal is to predict the mean value of sadness, anger, optimism, and joy in the next step.

C. Topic Multi-Perspective Analysis

Since experts may be interested in analysing topic relevance from different perspectives, the last tier allows a synergistic multi-perspective analysis of the topics extracted. For this purpose, the topic is analysed by means of the three parameters: sentiment, importance and spread in the community. The sentiment is represented by the sentiment score assessed on the topic for a specific emotional class, as it has been defined in Section B, the importance is based on the topic

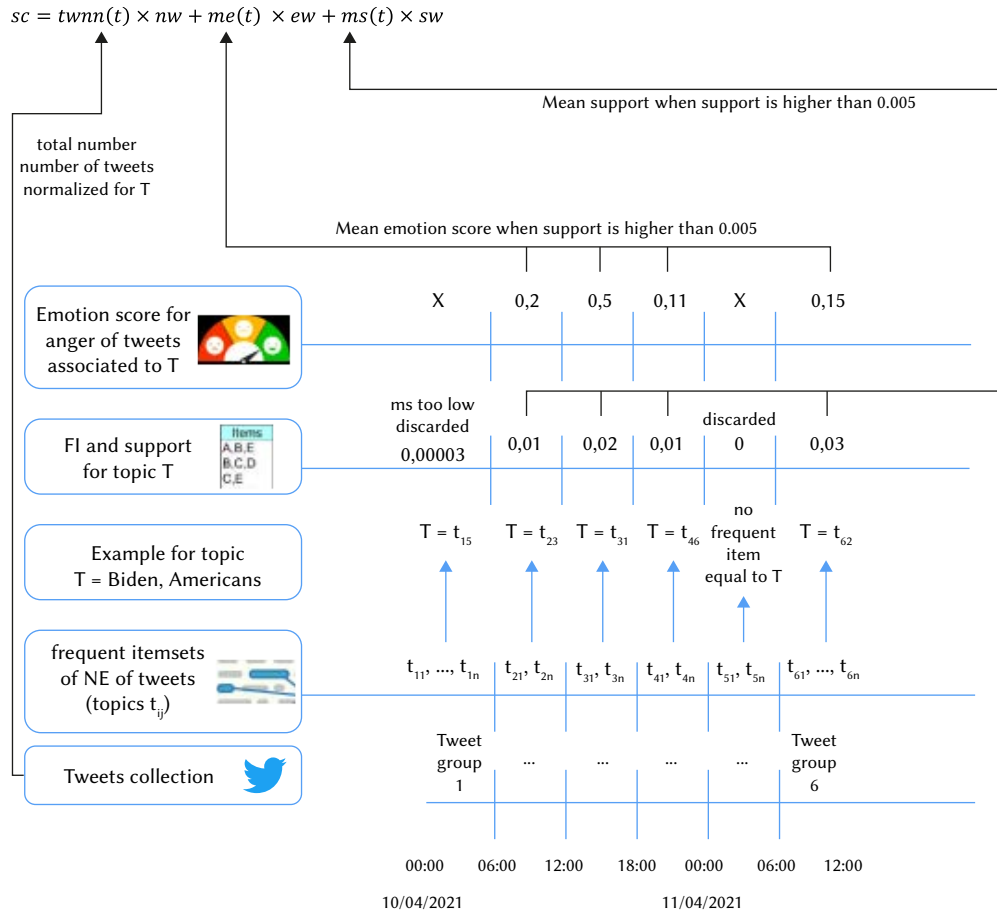


Fig. 3. Biden-americans: an example of the framework processing a topic.

support metric that represents the frequency of the topic over time compared to the other itemsets extracted from tweets (see Section A), and the topic spread evaluation is based on the number of topics in which the topic is present.

To provide a joint evaluation of topic relevance in the community, this tier employs a merged score function that combines the three above-mentioned parameters in a unique value depicting topic relevance. The merged score function is defined as a weighted function combining topic support, emotional class mean score, and the number of tweets in which the topic appears. Formally, let t be an itemset of terms, the merged score is assessed with the following weighted function:

$$sc = sw \cdot ms(t) + ew \cdot me(t) + nw \cdot twnn \quad (1)$$

where sw is the support weight, ms is mean support on all time-grids, ew is the weight associated to the average sentiment score, me is the average sentiment score over all time-grids, $twnn$ is the normalized number of tweets in which the t is present over all time-grids and nw is the weight associated to $twnn$. The average sentiment score (me) can be assessed for each of the four emotional classes considered by the framework.

The merged score is assessed for each topic among short- and long-term topics, hence a ranked list based on this score is generated to detect the most relevant short- and long-term topics among the extracted ones in the reference period. Moreover, experts can manually choose different weight values to generate and compare different views about the topic relevance. In fact, if experts want to focus on the emotional aspect, they can increase the weight associated with the mean emotional score (sw) and look out for those topics with the

highest emotional impacts, without ignoring the effective importance, and spread of the topic. The merged score allows a better dynamic analysis of topic relevance in the community, as it will be shown in the next section.

To summarize the whole framework functioning, Fig. 3 shows an example of running the framework in a time reference period displaying how the score is assessed for a specific topic T . The blue lines represent the time grid schema for data processing and the black lines show the stage at which the three merged function parameters are calculated or retrieved. In detail, first, the tweets are collected and arranged in groups in each time interval of the grid, hence the normalized number of tweets is assessed from data so as to be used in the merged score calculation. In the second stage, named entities are extracted from each group in the reference interval and arranged in frequent itemsets (FIs) representing the topics, thus the support for a topic T is calculated in each time interval allowing the calculation of the mean support (ms) that will be used for the topic T merged score calculation. Then, the sentiment score is calculated for an emotional class (anger in this example) so that the mean of the sentiment scores (me) for topic T over the intervals considered can be assessed and used for the final merged score calculation.

IV. A CASE STUDY ON LONG- AND SHORT-TERM TOPICS

To show the potential of the proposed framework, this section presents a real case scenario carried out by running the whole framework on tweets about the Covid-19 pandemic. For the sake of simplicity, we collected tweets with hashtags related to the Covid-19 general topic in the period going from 06/02/2021 to 14/02/2021. Then,

data have been processed by following the three-tier pipeline in Fig. 2 and the frequent itemsets of named entities representing topics extracted from tweets have been generated.

Since the main aim of the case study is to show how the proposed framework allows a comparative analysis of the effective short- and long-term topic relevance, the topic impact in terms of support score and sentiment score of every single topic extracted is evaluated through the merged metric (Eq. 1) that has been defined in Section C. Thanks to the merged score function, the framework returns two ranked lists of topics, one for the short-term and the other one for the long-term topics that have caused the strongest reactions among users and have been among the most discussed ones in the reference period.

The topics extracted by the framework have been ranked by using the merged score for the anger emotional class and shown in Tables I, II, and III for long-term topics and Tables IV, V and VI for short-term topics. Three different ranked lists are generated for each class of topic (i.e., short- and long- term topics) by setting different weight values for the three parameters, including anger emotion weight (ew), support weight (sw) and normalized number of tweets weight (nw).

Let us notice some difference between ranked lists of long- and short-term topics. For what concern the long-term ones, the highest-scored topic is *biden-americans* on most of the lists (see Tables II and III), which is a very general topic debated in 5,986 tweets, and to which users react with low anger (0.20). Since this topic is very general, it is not easy to associate it to news released in the reference period. However, the wide spread of the topic means that slight changes in negative emotions could influence a high number of people.

TABLE I. LONG-TERM TOPIC RANKED LIST FOCUSED ON SENTIMENT ($sw = 0.2$, $ew = 0.6$, $nw = 0.2$)

keys	sc	me	ms	$twnn$
republicans-rwpusa	0.545	0.815	0.014	0.266
senate-republicans	0.521	0.747	0.015	0.349
biden-americans	0.322	0.199	0.013	1.0
senate-eugene goodman fri-michaelart123	0.245	0.406	0.006	0.0
rt-rand paul	0.245	0.405	0.007	0.0

Legend: sc is the merged score, me is the average sentiment score over all time-grids, ms is the mean support over all time-grids, $twnn$ is the normalized number of tweets

TABLE II. LONG-TERM TOPIC RANKED LIST FOCUSED ON SUPPORT ($sw = 0.2$, $ew = 0.6$, $nw = 0.2$)

keys	sc	me	ms	$twnn$
biden-americans	0.248	0.199	0.013	1.0
senate-republicans	0.228	0.747	0.015	0.349
republicans-rwpusa	0.225	0.815	0.014	0.266
rt-rand paul	0.085	0.405	0.007	0.0
senate-eugene goodman fri-michaelart123	0.085	0.406	0.006	0.0

TABLE III. LONG-TERM TOPIC RANKED LIST FOCUSED ON THE NUMBER OF TWEETS IN WHICH THE TOPIC IS PRESENT ($sw = 0.2$, $ew = 0.2$, $nw = 0.6$)

keys	sc	me	ms	$twnn$
biden-americans	0.642	0.199	0.013	1.0
senate-republicans	0.362	0.747	0.015	0.349
republicans-rwpusa	0.326	0.815	0.014	0.266
senate-eugene goodman fri-michaelart123	0.082	0.406	0.006	0.0
rt-rand paul	0.082	0.405	0.007	0.0

TABLE IV. SHORT-TERM TOPIC RANKED LIST FOCUSED ON SENTIMENT ($sw = 0.2$, $ew = 0.6$, $nw = 0.2$)

keys	sc	me	ms	$twnn$
china-junta	0.634	0.955	0.018	0.29
asian-iamcindyachu-non-asian	0.631	0.914	0.006	0.406
2021-junta	0.628	0.953	0.024	0.259
china-2021	0.624	0.944	0.024	0.263
china-myanmar-2021	0.611	0.952	0.02	0.181
angelayarner-british	0.573	0.907	0.005	0.141
asian-asian americans	0.549	0.869	0.008	0.131
vp-asian	0.549	0.868	0.008	0.132
vp-asian americans	0.547	0.864	0.008	0.133
trump-covid	0.52	0.575	0.007	0.871
covid-billienomxtes	0.49	0.695	0.009	0.355
dwuhlfelderlaw-ron desantis	0.483	0.746	0.006	0.171
biden-florida	0.478	0.701	0.006	0.284
biden-trump	0.401	0.397	0.008	0.803
super bowl-dwuhlfelderlaw-tampa	0.377	0.585	0.008	0.125
super bowl-tampa	0.367	0.564	0.008	0.136
trump-gop	0.354	0.434	0.007	0.462
doctorpisspants-20s	0.353	0.444	0.011	0.422
trump-americans	0.344	0.401	0.008	0.511
0-daliagebrial	0.322	0.435	0.01	0.296
kylegriffin1-biden	0.303	0.295	0.013	0.616
senate-rand paul	0.3	0.436	0.008	0.184
senate-rt	0.293	0.42	0.008	0.196
a year ago today-gtconway3d	0.285	0.404	0.007	0.205
emekamba-nigeria	0.278	0.411	0.014	0.141

TABLE V. SHORT-TERM TOPIC RANKED LIST FOCUSED ON SUPPORT ($sw = 0.6$, $ew = 0.2$, $nw = 0.2$)

keys	sc	me	ms	$twnn$
trump-covid	0.293	0.575	0.007	0.871
iamcindyachu-asian	0.268	0.914	0.006	0.406
china-junta	0.26	0.955	0.018	0.29
2021-junta	0.257	0.953	0.024	0.259
china-2021	0.256	0.944	0.024	0.263
biden-trump	0.245	0.397	0.008	0.803
china-myanmar	0.238	0.952	0.02	0.181
covid-billienomxtes	0.216	0.695	0.009	0.355
angelayarner-british	0.213	0.907	0.005	0.141
asian-vp-asian americans	0.205	0.869	0.008	0.131
vp-asian	0.205	0.868	0.008	0.132
vp-asian americans	0.204	0.864	0.008	0.133
biden-florida	0.201	0.701	0.006	0.284
kylegriffin1-biden	0.19	0.295	0.013	0.616
trump-americans	0.187	0.401	0.008	0.511
dwuhlfelderlaw-ron desantis	0.187	0.746	0.006	0.171
trump-gop	0.184	0.434	0.007	0.462
doctorpisspants-20s	0.18	0.444	0.011	0.422
potus-america	0.178	0.161	0.008	0.706
0-daliagebrial	0.152	0.435	0.01	0.296
super bowl-dwuhlfelderlaw-tampa	0.147	0.585	0.008	0.125
super bowl-tampa	0.145	0.564	0.008	0.136
gop-americans	0.145	0.316	0.007	0.386
three weeks ago-america	0.135	0.206	0.01	0.439
supe-anildash	0.135	0.36	0.009	0.291

TABLE VI. SHORT-TERM TOPIC RANKED LIST FOCUSED ON SENTIMENT ($sw = 0.2$, $ew = 0.6$, $nw = 0.2$)

keys	sc	me	ms	twnn
trump-covid	0.639	0.575	0.007	0.871
biden-trump	0.563	0.397	0.008	0.803
potus-america	0.457	0.161	0.008	0.706
kylegriffin1-biden	0.431	0.295	0.013	0.616
asian-non-asian	0.428	0.914	0.006	0.406
trump-americans	0.388	0.401	0.008	0.511
china-junta	0.369	0.955	0.018	0.29
trump-gop	0.365	0.434	0.007	0.462
covid-billionomxtes	0.354	0.695	0.009	0.355
china-2021	0.351	0.944	0.024	0.263
2021-junta	0.35	0.953	0.024	0.259
doctorpisspants-20s	0.344	0.444	0.011	0.422
biden-florida	0.312	0.701	0.006	0.284
three weeks ago-america	0.307	0.206	0.01	0.439
myanma-2021	0.303	0.952	0.02	0.181
potus-covid	0.301	0.134	0.007	0.455
gop-americans	0.296	0.316	0.007	0.386
biden-gop	0.279	0.286	0.008	0.367
barbados-india	0.27	0.182	0.006	0.387
india-justintrudeau	0.269	0.141	0.007	0.4
angelarayner-british	0.267	0.907	0.005	0.141
0-daliagebrial	0.267	0.435	0.01	0.296
asian-asian americans	0.254	0.869	0.008	0.131
vp-asian americans	0.254	0.864	0.008	0.133
vp-asian	0.254	0.868	0.008	0.132

Now, let us pick the short-term topics with a higher merged score (0.63) in Table IV, which are about Myanmar junta, China and year 2021 (e.g., *china-junta*, *china-myanmar*, *china-myanmar-2021*, etc.). Even though these topics are discussed for a short time and do not have high mean support, Twitter users react to them by expressing considerable anger ($me=0.95$). Moreover, these topics identify specific news released in that period about the Myanmar military junta that took power with a coup d'état considering the Covid-19 pandemic restrictions as a violation imposed by the State Counsellor of Myanmar, and then killed and tortured hundreds of civilians, including children⁴.

The assignment of different weights to the merged score function contributes to build three different views on topics focusing on a specific concept but taking into consideration the others at the same time. This mechanism allows a comparative analysis that can indeed serve the detection of topics to pay attention to for avoiding destabilization effects:

1. Higher emotion weight: by giving higher weight to the sentiment score, the ranked lists (Tables I and IV) highlight those topics to which users react in the strongest way that may differ from the most-discussed topics (i.e., ranked lists focused on support and number of tweets). In fact, the long-term topic that caused the strongest users' emotional reactions is related to senate and republicans (see Table I), even though the most debated long-term topic is the already-mentioned *biden-americans* (see Tables II and III). The same goes for the short-term topic; the *china-junta-2021*, causing the strongest users' reactions (Table IV), is not discussed as much as the *trump-covid* and *biden-trump* which are the most debated topics (Tables V and VI).

2. Higher normalized tweet number weight: a normalized-tweet-number-focused ranked lists depict the importance of a topic. For instance, let us notice that *biden-trump* and *china-junta* are the topics with 80% and 29% of normalized number of tweets in the reference period (Table VI). Therefore, *biden-trump* has more importance than *china-junta*.
3. Higher support weight: the analysis mainly based on support gives a score to *biden-trump* that is less than the score of *china-junta*, meaning that, in the periods in which the topics are discussed, *china-junta* is more frequent than *biden-trump* and may be considered more important from this perspective (Table V). In other words, even if *china-junta* may be discussed within a smaller number of time-grid cells, during this time its mean frequency is considerably higher than the frequency of *biden-trump*.

The merged score allows to always consider all the parameters in a topic relevance evaluation even though a higher weight may be assigned to one of them. In fact, in the support-focused analysis (Table V), *china-junta* has higher ranking than *biden-trump* that does not only depend on support but also on the emotional score (i.e., *china-junta* causes definitely stronger emotional reactions than *biden-trump*).

V. EXPERIMENTATION

To test how much the proposed granular time-based framework is good for monitoring users' emotional reactions, a time-based test has been carried out to check out how long the framework has good performance at predicting users' reactions.

A. The Dataset

A Tweet dataset has been considered for tests: Coronavirus (COVID-19) Tweets Dataset [22]. This dataset is composed of CSV files including IDs and sentiment scores of the tweets related to the COVID-19 pandemic. The dataset includes 2,023,557,636 tweets in English, covering a global geographic area and a long period starting with the date of the first tweet on the topic which dates back to October 01, 2019. The real-time Twitter feed is monitored for coronavirus-related tweets using 90+ different keywords and hashtags that are commonly used while referencing the pandemic. For efficiency purposes, a subset of this dataset has been acquired, containing 2,000,000 tweets.

B. Methods

To check the feasibility of using our framework for users' reaction prediction, several state-of-the-art regression methods have been run and compared on the COVID-19 Tweets Dataset. Details on methods are reported below.

- **Linear regression** is a regression model consisting of a predictor variable and a dependent variable related linearly to each other.
- **Random Forest** is an ensemble learning method that averages the predictions made by multiple decision trees to perform regression.
- **Gradient boosting** is a predictive model combining an ensemble of weak prediction models for accomplishing regression.
- **K-nearest regressor** is a non-parametric method that approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood.

C. Metrics

Several metrics have been employed for tests:

- **MSE**. Mean Squared Error (**MSE**) refers to minimizing the mean squared error between predictions and expected values. It is calculated as the mean or average of the squared differences

⁴ <https://www.voanews.com/a/myanmar-junta-violations-may-amountto-crimes-against-humanity/6242469.html>

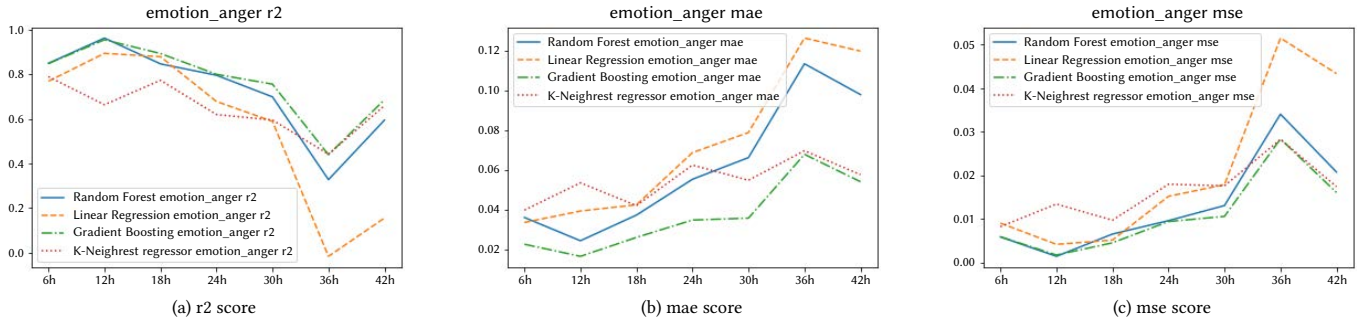


Fig. 4. Test results for the anger class.

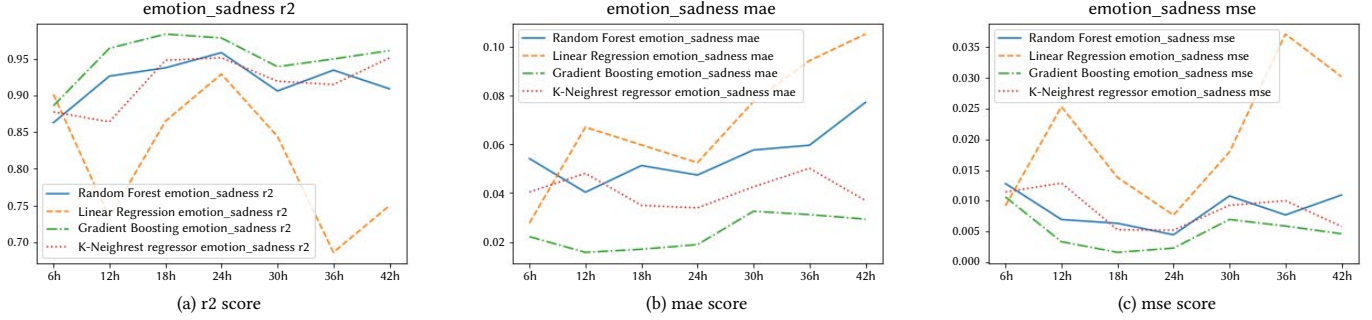


Fig. 5. Test results for the sadness class.

between predicted and expected target values in a dataset:

$$MSE = \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N} \quad (2)$$

where y_i is the i^{th} expected value in the dataset and \hat{y}_i is the i^{th} predicted value.

- **RMSE.** It is an extension of MSE that returns an error in the same unit of the target value. MSE allows to punish large errors by squaring the error, while RMSE reverses this operation through the square root:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}} \quad (3)$$

- **MAE.** Mean Absolute Error (MAE) does not give more or less weight to different types of errors, contrary to MSE and RMSE which punish larger errors more than smaller errors. MAE is calculated as the average of the absolute error values:

$$MAE = \frac{\sum_{i=1}^N (y_i - \hat{y}_i)}{N} \quad (4)$$

- **R^2 .** The coefficient of determination or R squared is statistics to evaluate how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. It is calculated by relating the residual sum of squares and the total sum of squares:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (5)$$

where $\bar{y} = \frac{\sum_{i=1}^N y_i}{N}$.

D. Test Results

The framework has been tested with each of the four methods introduced in Section B, then the results have been compared. The regressors have been applied to each of the four emotional classes. The selected dataset has been divided by 66% for training and 33% for test validation phases, hence all the regressors have been applied to predict the mean sentiment score in the next step by considering the mean

sentiment score in the previous four cells of the grid. Let us consider a prediction window as the number of cells after which the regressor provides users' sentiment prediction, the test has been designed as an incremental scheme that calculates the four metrics (Section C) on each regressor by incrementally increasing the prediction window by one at each test run. In other words, the regressor performance is evaluated firstly when it predicts the sentiment after one cell, then after two cells, then after three cells, and so on. The maximum number of cells to which regressor prediction accuracy is evaluated is 8. Since each interval is fixed to 6 hours, 8 cells correspond to 48 hours, therefore considering a scale of 8 cells, the regressor will be evaluated at predicting users' reactions after 6 h, 12 h, 18 h, 24 h, 30 h, 36 h, 42 h, 48 h. This time-based test allows to evaluate how long is the regressor good at predicting future users' emotional reactions.

From the figures, let us notice that, in general, Gradient boosting regressor outperforms all the other methods, followed by Random forest and k-nearest neighbor methods. Linear regression is definitely the method with the worst performance. The prediction accuracy over time changes with the emotional class, in detail, let us look Fig. 4c for the anger emotional class, r^2 score decreases after 18 hours, with the best performing method (gradient boosting) going down from 0.90 to 0.60, Random forest and k-nearest neighbor reaching 0.50 after the 30 hours, while Linear regression registers some negative steep dips. Regressors have definitely better performance on the other emotional classes, for sadness the r^2 score of the best-performing methods (e.g., Gradient boosting and Random forest) is constantly high lying in the range 0.9, 1s (Fig. 5c); for joy, r^2 score for three of the methods is high for 30 hours, after that values decrease a bit but lying in the range 0.6, 0.8s (Fig. 6a); for optimism, there is more variability, with the best-performing three regressors decreasing a bit after 12 hours but keeping high accuracy (r^2 score around 0.8) and going higher after 18 hours, but then having a steep dip after the 24 hours.

VI. CONCLUSION

The paper presented an emotion-aware framework for the analysis of long- and short-term topics over time and users' reaction to

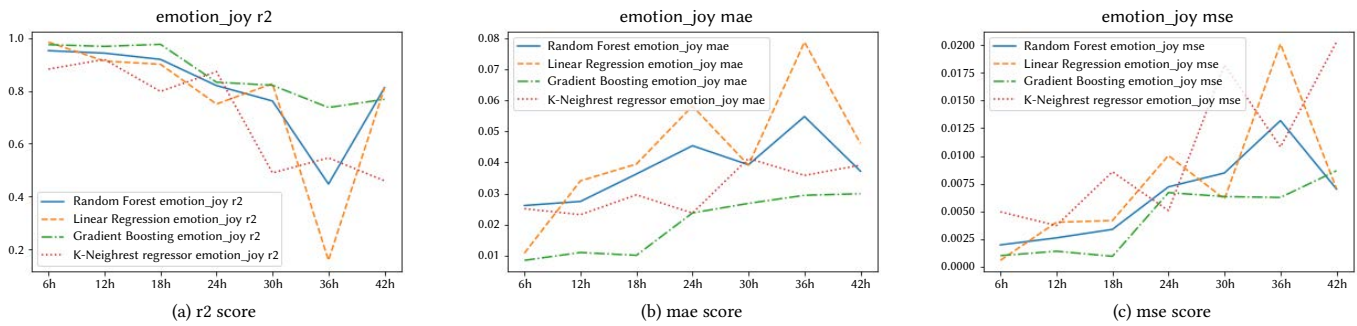


Fig. 6. Test results for the joy class.

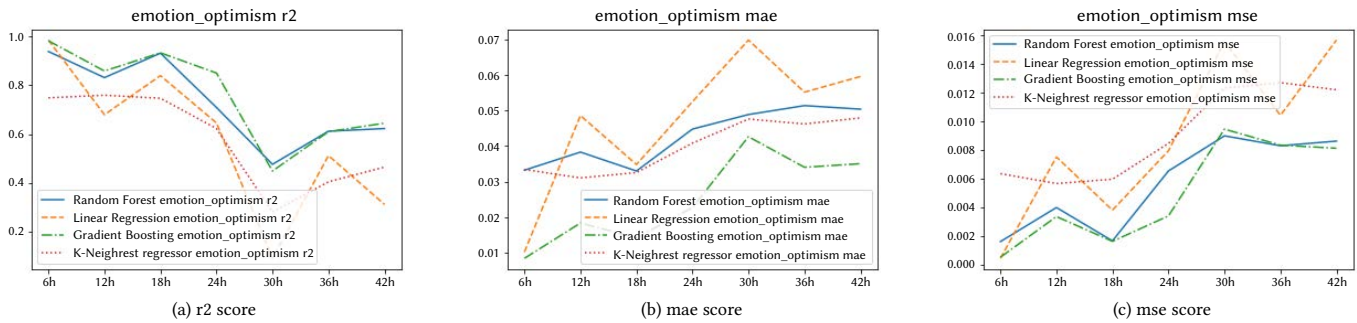


Fig. 7. Test results for the optimism class.

topics with the aim of supporting experts and institutions to keep disinformation on social networks under monitoring. This study introduced several important contributions to build monitoring systems for disinformation fighting in online social communities, including:

- A preliminary study showing the relevance of short-time-discussed topics in causing strong negative users' reactions.
- A time-grid-based approach to track topic frequency and emotional impact for the analysis of long- and short-time- debated topics.
- An emotion-aware topic modeling to support monitoring activities over time, including users' future reaction prediction.
- A score function combining topic frequency, sentiment and spread to support a robust multi-perspective topic relevance evaluation.

Future research intents will be focused on automatizing some processes, including an automatic weight assignment depending on the context (e.g., social, economical, political, etc.) and the analysis goal to find out the most relevant parameters for the topic relevance evaluation. Future research directions are also targeted at studying echo chamber effects in order to extend the developed short- and long-term topic detection model to help community analysers fight radicalization phenomena.

REFERENCES

- [1] E. Tonkin, "Chapter 2 - a day at work (with text): A brief introduction," in *Working with Text*, E. L. Tonkin, G. J. Tourte Eds., Chandos Information Professional Series, Chandos Publishing, 2016, pp. 23–60, doi: <https://doi.org/10.1016/B978-1-84334-749-1.00002-0>.
- [2] X. Zhang, A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Information Processing & Management*, vol. 57, no. 2, p. 102025, 2020, doi: <https://doi.org/10.1016/j.ipm.2019.03.004>.
- [3] S. Loomba, A. Figueiredo, S. Piatek, K. de Graaf, H. Larson, "Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa," *Nature Human Behaviour*, vol. 5, pp. 1–12, 2021, doi: 10.1038/s41562-021-01056-1.
- [4] Y. Kirill, I. G. Mihail, M. Sanzhar, M. Rustam, F. Olga, M. Ravil, "Propaganda identification using topic modelling," *Procedia Computer Science*, 9th International Young Scientists Conference in Computational Science, YSC2020, 05-12 September 2020, vol. 178, pp. 205–212, 2020, doi: <https://doi.org/10.1016/j.procs.2020.11.022>.
- [5] J. Dao, J. Wang, X. Zhang, "YNU-HPCC at SemEval- 2020 task 11: LSTM network for detection of propaganda techniques in news articles," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online), Dec. 2020, pp. 1509– 1515, International Committee for Computational Linguistics.
- [6] V. Ermurachi, D. Gifu, "UAIC1860 at SemEval- 2020 task 11: Detection of propaganda techniques in news articles," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online), Dec. 2020, pp. 1835–1840, International Committee for Computational Linguistics.
- [7] V.-A. Oliinyk, V. Vysotska, Y. Burov, K. Mykich, V. B. Fernandes, "Propaganda detection in text data based on nlp and machine learning," in *MoMLet+DS*, 2020.
- [8] E. Ferrara, "Contagion dynamics of extremist propaganda in social networks," *Information Sciences*, vol. 418-419, pp. 1–12, 2017, doi: <https://doi.org/10.1016/j.ins.2017.07.030>.
- [9] S. Khan, S. Hakak, N. Deepa, B. Prabadevi, K. Dev, S. Trelova, "Detecting covid-19-related fake news using feature extraction," *Frontiers in Public Health*, vol. 9, 2022, doi: 10.3389/fpubh.2021.788074.
- [10] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, J. Vilares, "Sentiment analysis for fake news detection," *Electronics*, vol. 10, no. 11, 2021, doi: 10.3390/electronics10111348.
- [11] M. Huddar, S. Sannakki, V. Rajpurohit, "Attentionbased multi-modal sentiment analysis and emotion detection in conversation using RNN," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 112-121, 2021, doi: 10.9781/ijimai.2020.07.004.
- [12] C.-H. Chen, P.-Y. Chen, J. Lin, "An ensemble classifier for stock trend prediction using sentence-level chinese news sentiment and technical indicators," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 3, pp. 53–64, 2022, doi: 10.9781/ijimai.2022.02.004.
- [13] M. Charquero-Ballester, J. G. Walter, I. A. Nissen, A. Bechmann, "Different types of covid- 19 misinformation have different emotional valence on twitter," *Big Data & Society*, vol. 8, no. 2, p. 20539517211041279, 2021, doi: 10.1177/20539517211041279.
- [14] Y. Wang, R. Han, T. Lehman, Q. Lv, S. Mishra, "Analyzing behavioral changes of twitter users after exposure to misinformation," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21, New York, NY, USA, 2021, p. 591–598, Association for Computing Machinery.
- [15] N. Kalantari, D. Liao, V. G. Motti, "Characterizing the online discourse

in twitter: Users' reaction to misinformation around covid-19 in twitter," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 4371–4380.

- [16] S.-Y. Lin, Y.-C. Kung, F.-Y. Leu, "Predictive intelligence in harmful news identification by bert-based ensemble learning model with text sentiment analysis," *Information Processing & Management*, vol. 59, no. 2, p. 102872, 2022, doi: <https://doi.org/10.1016/j.ipm.2022.102872>.
- [17] S. Kwon, A. Park, "Understanding user responses to the covid-19 pandemic on twitter from a terror management theory perspective: Cultural differences among the us, uk and india," *Computers in Human Behavior*, vol. 128, p. 107087, 2022, doi: <https://doi.org/10.1016/j.chb.2021.107087>.
- [18] P. Koukaras, C. Tjortjjs, D. Rousidis, "Mining association rules from covid-19 related twitter data to discover word patterns, topics and inferences," *Information Systems*, vol. 109, p. 102054, 2022, doi: <https://doi.org/10.1016/j.is.2022.102054>.
- [19] K. Garcia, L. Berton, "Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa," *Applied Soft Computing*, vol. 101, p. 107057, 2021, doi: <https://doi.org/10.1016/j.asoc.2020.107057>.
- [20] J. Han, J. Pei, Y. Yin, "Mining frequent patterns without candidate generation," *SIGMOD Rec.*, vol. 29, p. 1–12, may 2000, doi: 10.1145/335191.335372.
- [21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *ArXiv*, vol. abs/1907.11692, 2019.
- [22] R. Lamsal, "Coronavirus (covid-19) tweets dataset," 2020. [Online]. Available: <https://dx.doi.org/10.21227/781w-ef42>, doi: 10.21227/781w-ef42.



Francesco David Nota

Francesco David Nota (Member, IEEE) received the master's degree in Business Innovation and Informatics from the University of Salerno, Italy, in 2020. He is currently pursuing the Ph.D. degree in innovation sciences for defence and security–digital transformation and cybersecurity with the Center for Higher Defence Studies (CASD). His research interests include Cognitive Warfare.



Danilo Cavaliere

Danilo Cavaliere received both the master degree cum laude in computer science and the Ph.D. degree from the University of Salerno (Italy) in 2014 and 2020, respectively. He works now as postdoctoral researcher for the same institution. Dr. Cavaliere is in the editorial board of *Neurocomputing* journal and in the Program Committee for IEEE Symposium on Intelligent Agents, in

the conference IEEE SSCI since 2019. His research interests include artificial and computational intelligence, data science, intelligent agents, data mining and knowledge discovery, scientific fields on which he has published many papers.



Giuseppe Fenza

Giuseppe Fenza (Member, IEEE) received the degree and Ph.D. degrees in computer sciences from the University of Salerno, Italy, in 2004 and 2009, respectively. He is currently an Associate Professor of computer science at the University of Salerno. The research activity concerns computational intelligence methods to support semantic-enabled solutions and decision-making. He has over 60

publications in fuzzy decision making, knowledge extraction and management, situation and context awareness, semantic information retrieval, service oriented architecture, and ontology learning. More recently, he worked in automating open source intelligence and big data analytics for counterfeiting extremism and supporting information disorder awareness.



Vincenzo Loia

Vincenzo Loia (Senior Member, IEEE) received the degree in computer science from the University of Salerno, Italy, in 1985, and the Ph.D. degree in computer science from the Université Pierre Marie Curie Paris VI, France, in 1989. He is currently a Computer Science Full Professor at the University of Salerno, where he worked as a Researcher, from 1989 to 2000, and as an Associate Professor, from

2000 to 2004. He is the Co-Editor-in-Chief of *Soft Computing* and the Editor-in-Chief of *Journal of Ambient Intelligence and Humanized Computing*. He serves as an editor for 14 other international journals.

Tourism-Related Placeness Feature Extraction From Social Media Data Using Machine Learning Models

P. Muñoz¹, E. Doñaque², A. Larrañaga³, J. Martínez^{4*}, A. Mejías⁵

¹ Department of Financial and Accounts Economics, University of Vigo, Pontevedra (Spain)

² Possible Incorporated S. L.

³ CINTECX, University of Vigo, Pontevedra (Spain)

⁴ Department of Applied Mathematics, University of Vigo, Pontevedra (Spain)

⁵ Department of Business Organization and Marketing, University of Vigo, Pontevedra (Spain)

Received 1 February 2022 | Accepted 2 December 2022 | Published 21 December 2022



ABSTRACT

The study of *placeness* has been focus for researchers trying to understand the impact of locations on their surroundings and tourism, the loss of it by globalization and modernization and its effect on tourism, or the characterization of the activities that take place in them. Identifying places that have a high level of placeness can become very valuable when studying social trends and mobility in relation to the space in which the study takes place. Moreover, places can be enriched with dimensions such as the demographics of the individuals visiting such places and the activities they carry in them thanks to social media and modern machine learning and data mining methods. Such information can prove to be useful in fields such as urban planning or tourism as a base for analysis and decision-making or the discovery of new social hotspots or sites rich in cultural heritage. This manuscript will focus on the methodology to obtain such information, for which data from Instagram is used to feed a set of classification models that will mine demographics from the users based on graphic and textual data from their profiles, gain insight on what they were doing in each of their posts and try to classify that information into any of the categories discovered in this article. The goal of this methodology is to obtain, from social media data, characteristics of visitors to locations as a discovery tool for the tourism industry.

KEYWORDS

M3 Inference, Machine Learning, Social Media, Tourism, Word2Vec.

DOI: 10.9781/ijimai.2022.12.003

I. INTRODUCTION

TOURISM is a source of wealth and sustained growth. The relationship between tourism and economic growth has been explained in several studies [1], [2]. Until 2019, the main motivation for international travel was tourism, being the reason for 56% of them, followed by visiting friends and family, health, religion and other purposes (27%), and business travel (13%). Until that time, tourism was the world's third largest export, after chemicals and fuels [3]. International tourism spending experienced an average annual growth of 4.6% between 2010 and 2018 [4], (2020). Even in 2019, this sector grew by 3.5%, contributing 10.3% to global GDP (i.e., Gross Domestic Product) and 28.3% to global exports of services [5]. Tourism has also undergone a qualitative transformation. The externalities associated with mass tourism, the emergence of a great diversity of tourism products to face competition from destinations, the change in the profile of the tourist who seeks different experiences focused on culture, nature, authenticity, among others [6], [7] have been some of the factors associated with this transformation. Tourism activities related to the cultural heritage-nature binomial the backbone of cultural tourism [8].

The concept of cultural heritage (i.e., CH) is very broad, including elements such as landscapes, historical sites, works of art, biodiversity, traditions, social values, sensory experiences, among others. In its contemporary meaning, it is made up of a tangible or physical component, the tangible cultural heritage; and an intangible component, the non-material cultural heritage [9]–[11]. The valuation of CH has been imposed in the narrative that guides the design of tourism products, due to the great dynamizing potential of the society and economy of the territories where it is promoted [12]. On the other hand, it has favored the development of cultural heritage tourism, as a tourism with a global dimension [13].

However, a tourist location may succeed to the point of generating a problem, of discomfort of residents [14] or disturbance of the environment [15]. Furthermore, the location of the tourist destination is dynamic [16]. In the context of globalization and increased human mobility, placeness can be affected by geopolitical issues, wars, terrorism, security threats and health emergencies among others [17], [18], [19].

In this sense, the study of the factors that influence the formation of placeness of tourism products is increasingly being investigated [20], [21]. *Placeness* is defined as the uniqueness of a place determined by the set of its natural and historical features, its tangible and intangible cultural assets, emotions and sensations that the place can generate both to the inhabitants of the destination and tourists [22],[23]. Social networks such as Instagram are a valuable source of data for the study of the aforementioned factors that influence the creation of *placeness*.

* Corresponding author.

E-mail address: javmartinez@uvigo.es

Thus, drawing on previous work that defined the concept, the availability of social media data and machine learning and data mining techniques, we intend study the viability of extracting *placeness* features in accordance to the ontological approach proposed by [24]. We expose here a case study that focuses on the demographic and activity information in a clearly defined area and period of time. We intend to show-case a small scale study with the goal of introducing the methods as a first stage, but we aim to extend this to larger studies covering several locations in a wide area as a mean to discover new potential touristic opportunities.

A. Related Works

Previous research groups have used social media to infer information about people or places and have developed tools and methodologies which provide promising results. Noe, the two key articles on which our work is based are described [25], [26], [27], [28].

- Inferring *placeness* from Starbucks In [24] the authors focus their research around *placeness* as "the sense of place" and its importance in architecture or urban design. In their study they define an ontology for placeness which describes it as a relation to four factors: place, visitors, time and activity. They came up with a novel methodology to extract *placeness* features from Instagram posts and characterize a specific location given the information inferred for the given factors. They show-cased an study conducted from posts tagged in Starbucks in three major cities and compared its results to the current big data based approach and showed promising results from the relatively small amount of data they dealt with.
- Inferring *demographics* from social media A deep learning system is specified in [29] as an alternative to infer demographic data from social media users while providing resilience against biases that favour dominant languages and groups. The results achieved in their study were very promising, reaching accuracies higher than other demographic inference systems while providing better support for text in different languages, supporting up to 32 languages.

II. METHODOLOGY AND MATHEMATICAL BACKGROUND

Drawing on the works presented, a particular inference pipeline was set up to enrich Instagram posts with the same dimensions defined in [24] with the support of the M3 Inference pre-trained model to deal with cases where the images do not offer demographic information based on face recognition [30] [31]. On top of that, an education estimator was added based on a linguistic formula to measure the readability score of posts [32].

A. M3 Inference

According to Z. Wang et. al [29], this deep learning system is named after its multimodal, multilanguage and multi-attribute inference capabilities because it has been designed and trained integrating different machine learning models to extract features closer to the input, to support 32 languages and infer three different attributes. Fig. 1 shows the input data the model requires and the output is obtained from it.

This system's architecture combines DenseNet [33] [34] for image classification, and a 2-stack bidirectional character-level Long Short-Term Memories [35] neural networks as input models that are later combined in what they call a *dropout layer* followed by, in sequential order, two densely connected layers, one Rectified Linear Unit activation layer and three different output layers that apply a SoftMax [36] function to the output in order to represent the probabilistic score of each category. This system has an added advantage of resilience against missing data thanks to its dropout layer, which is trained to

work when data from a source may no be available or reliable. For example, in cases where the image model cannot decide on any category, the dropout layer would give more weight to the input from the text based models.

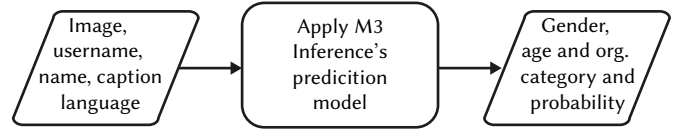


Fig. 1. M3 Inference's input and output.

B. Flesch Reading Ease

The *Flesch Reading Ease* is a simple and effective metric to measure readability invented in 1948 by Rudolph Flesch, which can be adapted to several languages [32]. Such expression (equation 1) gives a score related to the difficulty of the text analyzed, drawing a value within the range of 0 to 100 as a result. The lower readability score, the higher the level of education that the transmitter is supposed to have.

(1)

Table I shows the distribution of education levels according to this indicator. Therefore, a text made with short sentences and simple words would have a very high score, and a text made up of long and complex sentences would have a small value assigned.

TABLE I. FLESH READING EASE SCORE DISTRIBUTION

Flesch Reading Ease Categories	
Score	Level of difficulty
90-100	5th grade - Very easy
80-90	6th grade - Easy
70-80	7th grade - Fairly easy
60-70	8th to 9th grade - Plain English
50-60	10h to 12th grade - Fairly difficult
30-50	College - Difficult
10-30	College graduate - Very difficult
0-10	Professional - Extremely difficult

C. Word2Vec Model

In the area of natural language processing, words are generally interpreted as associated symbols that do not necessarily have to have meaning; thus, the word *tourism* could have an associated id such as *id1*. The problem of following this nomenclature lies precisely in not keeping a relationship between similar terms, and it is for this reason that space vector models have arisen, pretending to group words that belong to the same lexical family.

From this idea was born the Word2Vec model, which is mainly found in two forms: Continuous Bag-of-Words (CBOW) or Skip-Gram. The former consists of trying to determine the word in the middle of a text from the context taken from the rest surrounding the word of interest, and is the one used for the present work. On the other hand, the latter is based on trying to predict context of the text based on a given word of interest.

The way to do this is by using a technique called one-hot vector, which is based on creating a vector with as many zeros as words in the text and assigning a 1 to the position where the word of interest is located. Thus, for instance, if you have the sentence *Today is sunny* and you want to encode the word *sun*, the vector would look like [001].

The output of the Word2Vec Model is a vocabulary in which each item has an associated vector that can be used to identify similarities and relationships between all the words that characterize the images

from instagram, which in this case is composed of a 300-dimensional Word2vex model.

D. Principal Component Analysis

After completing the word encoding, it was decided to perform a step to reduce the dimensionality of the data from 300 to 2-dimensional, in order to both facilitate and optimize its subsequent classification. It is important to note that technically one dimension would have been sufficient to explain the variability of the vectors (since the first component is able to explain 90% of the variability); however, two have been chosen in order to follow the recommended methodology and provide greater representation and explanation of the differences in the data. *Principal Component Analysis*, i.e. PCA, is a reduction dimensionality technique used to improve the understanding of a large dataset, minimizing the information loss by creating new variables that are not correlated [37] [38]. In this way, enough components have been retained to create a two-dimensional space, which explains or contains more than 90% of the variance of the data.

E. Unsupervised Clustering Algorithm: KMeans

Once the reduction of the vector space corresponding to the tags that were taken from the instagram images is completed, the next step consists of applying one unsupervised algorithms known as KMeans. The term unsupervised implies that the input data does not contain labels, so the clustering is performed by similarity in the vectors representing the chosen words.

A parameter k is defined, which refers to the number of centroids to be searched for in the dataset, i.e. the number of clusters to be obtained. These centroids correspond to the center of the cluster in question. Once the number of clusters is selected, the algorithm starts iterating to optimize the position of the centroids of each cluster. In such a way, it will only stop if the centroid position stabilizes after a number of iterations, or if a maximum number of iterations is reached.

1. Silhouette Score

Aiming to identify the optimal number of clusters to be chosen for the KMeans clustering algorithm, the average silhouette method is evaluated in order to determine how well each item is classified within its cluster. The higher the average silhouette width, the more appropriate the item is classified. Calculations for different numbers of groupings show that the optimum is two, since it corresponds to a value of 0.76 according to the silhouette method, being the highest of all the tests performed (Table II).

TABLE II. SILHOUETTE METHOD COEFFICIENTS

Nº clusters	2018	2021
2	0.7576	0.7699
3	0.6696	0.6781
4	0.6392	0.6484
5	0.6124	0.6268
6	0.5729	0.5640
7	0.5707	0.5658

III. RESULTS

This section describes the data used in this study, discusses the challenges in collecting it and explains its format as a preamble. The enrichment process and its components are described, what builds up the inference pipeline, and the different machine learning approaches used are discussed - directly or indirectly - as well as their strengths. An extra section will describe the activity classification model that has been developed for this study, explaining the different stages and techniques required to build it.

A. Data Description and Enrichment

The data used in this study is a set of 10,000 Instagram posts tagged in Vigo, Spain [39] and comprises two different datasets: 5,000 posts from the last two weeks of May 2018 and 5,000 posts from the last two weeks of May 2021. The goal of dealing with these datasets is to try and find differences in visitors - namely, people posting at a given location - before an after an event expected to have impacted travellers' and visitors' behaviour, such as the SARS-CoV-19 pandemic. It is important to note that, while desirable, obtaining large amounts of posts from Instagram is not an easy feat, thus efforts to provide inference tools from low amounts of data are required.

Table III shows an example of the data used to conduct the study in table format. More information could be retrieved, but these were the only parameters our enrichments needed.

TABLE III. DATA FORMAT EXAMPLE

image_url	https://instagram...
caption	Enjoy these moments...
username	the_foobaz
full_name	Eugenio Doe

Our data processing pipeline is shown in Fig. 2 as a set of independent processes that infer the different dimensions of interest. This process is referred as the enrichment of the data set because it infers valuable information from *raw* data. The overall pipeline consists of a wide range of machine learning and data mining techniques combined with the usage of external MLaaS1 platforms that allow us to get some promising results with small sets of data.

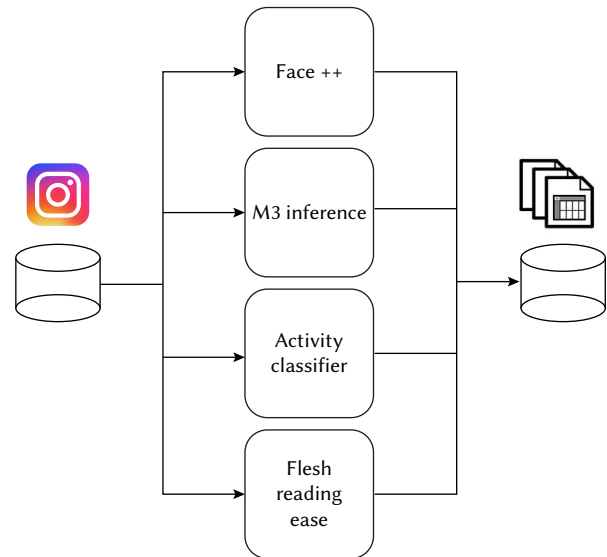


Fig. 2. Enrichment on data.

To obtain demographic information two machine learning methods are utilized: Face++ platform [30] and M3Inference [29]. The first uses facial recognition to identify faces and estimates their demographics based on them, while the second analyses the image and text data from the user to make the same inference. As explained before, the latter is used to obtain information where the former fails to recognize any face. The activity dimension is also inferred using machine learning and data mining techniques described later.

A different approach, however, is taken for the educational dimension, where instead of relying machine learning models the Flesch Readability formula was used. One such approach became the seed for more sophisticated evaluations, and even adaptations

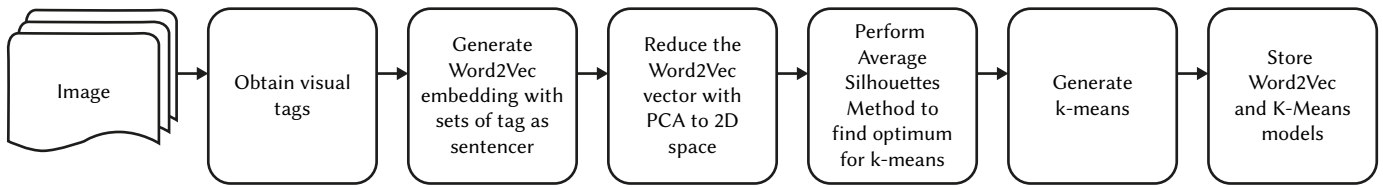


Fig. 3. Activity classifier generation process.

in different languages. Moreover, it offered flexibility on multiple languages -making switching between them straight forward thanks to the python implementation found in [40].

Some of these systems are closed source, such as Face++ and Azure's Cognitive Services [41] so no details on their techniques can be discussed, but provided their services cover the areas of facial and object recognition, it can be assumed there's a high degree of deep or convolutional neural networks at work given their outstanding performance on image classification tasks [42]. That is the case for one of the ways the demographic dimension of a subset of posts is obtained from the data set through Face++'s face detection web API, which searches for faces in images, analyses them and infers their age and gender, among other things, and returns that to the caller.

B. Clustering

1. Activity Classifier. Model Generation

Based on the work of [24] we made our own implementation of the activity classifier with the data we had. The overall process of generating the classifier is shown in Fig. 3, which mentions the techniques used in this approach.

The main parts of the process are explained as such:

1. Obtain visual tags from each image with Azure's computer vision API [41] and make sentences by joining all of the tags into text. These tags can be anything that the API recognizes, such as: outdoors, beach, water, sand, bikini.
2. Train a *Word2Vec* [43], [44] with the sentences to generate a word embedding that learns relationships between words. For example, words that repeatedly appear in the same sentences together will be more similar than those that do not.
3. Reduce the vector space from the *Word2Vec* model using PCA [38] to project the two main components - which provided the highest variance - into a new 2D space in order to allow for more efficient clustering. Other approaches were tested, such as UMAP [45] and T-SNE [46], but after comparing the final clusters by means of image sampling and by the sets of words that defined them, it was concluded that PCA offered similar results in significantly less time. On T-SNE, it was decided not to be a good option to make a reusable model because it is not deterministic, therefore there is no guarantee that new similar data would be classified into the right clusters.
4. Compute the Silhouette Score [47] for different values of k to discover the most efficient value for the K-Means clustering algorithm.

2. Data Classification Interpretation

Fig. 4 shows the comparison between the age profiles found within each grouping. First of all, it is noteworthy that the first cluster contains a greater number of individuals than the second cluster, which is composed mostly of a female profile. The age focus is centered between 25-30 years for both clusters, showing a downward trend towards older ages and no difference pre- and post-pandemic. On the other hand, for the second of the clusters a greater comparison between the profiles can be seen, counting a lower presence of women

for 2021. The center of age shifts more towards 30 years of age and the downward trend seen in the first cluster is lost.

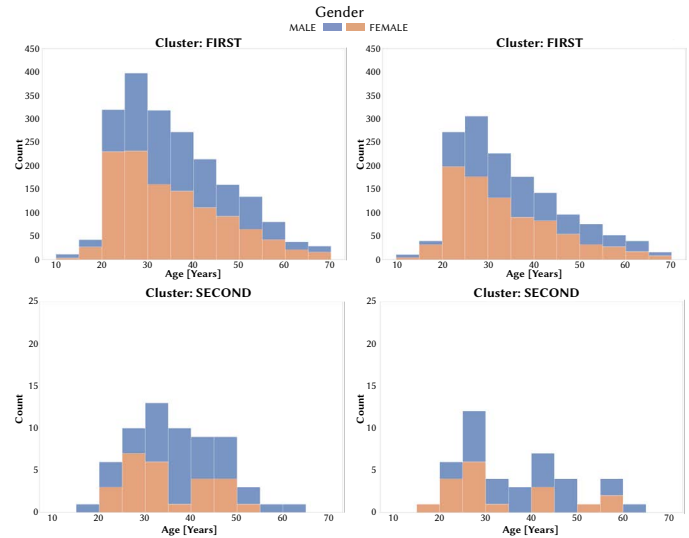


Fig. 4. Age comparison between clusters from 2018 (left) and 2021 (right)..

Moreover, Fig. 5 shows a box plot for each year, comparing the difference in the education profile of the publications between the first and second clusters, taking into account gender. In the first of the clusters, a higher level of Flesch's pre-pandemic indicator in men can be seen, which is equivalent to a lower difficulty of the analyzed text. This tendency is also seen to a lesser extent in the case of women, who present a greater dispersion of the data and an average that remains around 50 in 2018, decreasing towards 40 by 2021. In the case of the second cluster, it is clear that for both women and men, the level of difficulty of the texts analyzed increased post-pandemic.

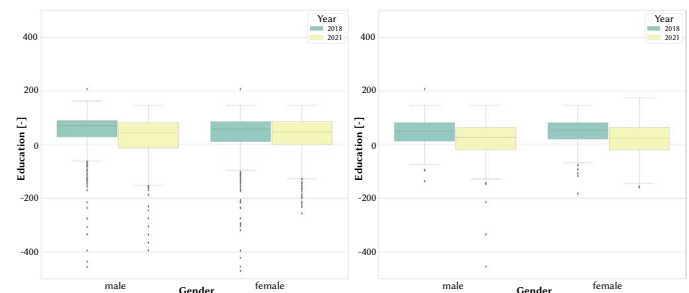


Fig. 5. Age comparison between years from first (left) and second (right) clusters.

IV. CONCLUSIONS AND FUTURE LINES

This work has developed a methodology for analysing tourism through the prism of complex mathematical algorithms, based on unstructured data extracted from social networks.

As a conclusion of this work, it is possible to appreciate the great potential offered by the analysis of information from social networks for the identification of tourism profiles. Serving in this case to locate differences between people visiting the city of Vigo between the years 2018 and 2021. In the results, a more notable difference has been observed in the education levels of the profiles, rather than in the age ranges; being logical that most of them focus on lower ages where social networks are more successful.

Beyond that, the geolocated image extraction and analysis methodology implemented in the present work has great potential for comparisons on larger amounts of data and even between tourism profiles between cities.

Once the results obtained in this work have been analysed, new algorithms are being developed that integrate several data sources, as well as the improvement of enrichment techniques, always oriented towards decision-making in the tourism sector.

ACKNOWLEDGMENT

Pilar Muñoz has received support from the Spanish ministry of Science and Research (grant PID2020-116040RB-I00). The work of Ana Larrañaga has been supported by the 2020 predoctoral grant of the University of Vigo.

REFERENCES

- [1] J. G. Brida, S. London, M. Rojas, "El turismo como fuente de crecimiento económico: impacto de las preferencias intertemporales de los agentes," *Investigación económica*, vol. 73, pp. 59 – 77, 09 2014.
- [2] I. Cortés-Jiménez, "Which type of tourism matters to the regional economic growth? the cases of Spain and Italy," *International Journal of Tourism Research*, vol. 10, no. 2, pp. 127–139, 2008, doi: <https://doi.org/10.1002/jtr.646>.
- [3] W. T. Organization, "International tourism highlights," 2019.
- [4] U. N. W. T. Organization, "Unwto global tourism dashboard. country profile - outbound," 2020.
- [5] W. Travel, T. Council, "Research – economic impact reports.," 2020.
- [6] A. Santana Talavera, "Patrimonios culturales y turistas: unos leen lo que otros miran," *PASOS : Revista de Turismo y Patrimonio Cultural*, vol. 1, 01 2003, doi: [10.25145/j.pasos.2003.01.001](https://doi.org/10.25145/j.pasos.2003.01.001).
- [7] F. Jiménez, C. y Seo, "Patrimonio cultural inmaterial de la humanidad y turismo.," *International Journal of Scientific Management and Tourism*, vol. 4, no. 2, pp. 349– 366, 2018.
- [8] G. Yudice, "El recurso de la cultura.," *Gedisa. Barcelona*, 2001.
- [9] UNESCO, "Convención para la salvaguarda del patrimonio cultural inmaterial de la unesco.," 2003.
- [10] J. Arévalo, "La tradición, el patrimonio y la identidad," pp. 925–955, 2004.
- [11] M. Timón Tiemblo, M.P. y Domingo Fominaya, "Resumen del plan nacional de salvaguarda del patrimonio cultural inmaterial," *Anales del Museo Nacional de Antropología*, vol. 14, pp. 29–44.
- [12] J. Nared, D. Bole, *Participatory Research on Heritage- and Culture-Based Development: A Perspective from South-East Europe*, pp. 107–119. Cham: Springer International Publishing, 2020.
- [13] B. A. Adie, C. M. Hall, "Who visits world heritage? a comparative analysis of three cultural sites," *Journal of Heritage Tourism*, vol. 12, no. 1, pp. 67–80, 2017, doi: [10.1080/1743873X.2016.1151429](https://doi.org/10.1080/1743873X.2016.1151429).
- [14] C. Milano, M. Novelli, J. M. Cheer, "Overtourism and tourismphobia: A journey through four decades of tourism development, planning and local concerns," *Tourism Planning & Development*, vol. 16, no. 4, pp. 353–357, 2019, doi: [10.1080/21568316.2019.1599604](https://doi.org/10.1080/21568316.2019.1599604).
- [15] P. L. Winter, S. Selin, L. Cervený, K. Bricker, "Outdoor recreation, nature-based tourism, and sustainability," *Sustainability*, vol. 12, no. 1, 2020, doi: [10.3390/su12010081](https://doi.org/10.3390/su12010081).
- [16] Y. Deng, C. Li, "Research progress, theories review and trend forecast on placeness of tourism destination," in *Proceedings of the 3rd International Seminar on Education Innovation and Economic Management (SEIEM 2018)*, 2019/01, pp. 431–434, Atlantis Press.
- [17] R. E., "Classics in human geography revisited, place and placelessness.," *Progress in Human Geography*, vol. 24, no. 4, p. 613, 2000.
- [18] S. A. Bowen, D. R. E., "Tourist satisfaction and beyond: tourist migrants in mallorca.," *International journal of tourism research*, vol. 10, no. 2, pp. 141–153, 2008.
- [19] A. Bowen, "War-affected children in three african short stories: Finding sanctuary within the space of placelessness.," *Commonwealth Essays and Studies*, vol. 42, no. 2, 2020.
- [20] T. Wenyue, "The influence and significance of tourism development on placeness.," *Tourism Tribune*, vol. 28, no. 4, pp. 9–11, 2013.
- [21] L. Leilei, "The spatial cognition process and law of tourist destination image.," *Scientia Geographica Sinica*, vol. 6, pp. 563–568, 2000.
- [22] W. B., "Regional tourism planning principles.," *China Travel and Tourism Press*, 2001.
- [23] K. X. Zhou S Y, Yang H Y, "The structuralistic and humanistic mechanism of placeness: A case study of 798 and m50 art districts.," *Geographical Research*, vol. 30, no. 9, pp. 1566–1576, 2011.
- [24] G. Kalra, M. Yu, D. Lee, M. Cha, D. Kim, "Ballparking the urban placeness: A case study of analyzing starbucks posts on instagram," in *International Conference on Social Informatics*, 2018, pp. 291–307, Springer.
- [25] M. K. . M. F. Pfeffer, J., "War-affected children in three african short stories: Finding sanctuary within the space of placelessness.," *Commonwealth Essays and Studies*, vol. 7, no. 1, p. 50, 2018.
- [26] B. E. Rossi, L., A. Torsello, "Venice through the lens of instagram: A visual narrative of tourism in Venice.," *Companion Proceedings of the The Web Conference 2018*, pp. 1190–1197, 2018.
- [27] K. Jang, Y. Kim, "Crowd-sourced cognitive mapping: A new way of displaying people's cognitive perception of urban space.," *PLoS ONE*, vol. 14, no. 6, p. e0218590, 2019.
- [28] U. S. Hasan S, Zhan X, "Understanding urban human activity and mobility patterns using large-scale location-based data from online social media.," *PLoS ONE*, vol. 14, no. 6, p. e0218590, 2003.
- [29] Z. Wang, S. Hale, D. I. Adelani, P. Grabowicz, T. Hartman, F. Flöck, D. Jurgens, "Demographic inference and representative population estimates from multilingual social media data," in *The world wide web conference*, 2019, pp. 2056–2067.
- [30] Beijing Kuangshi Technology Co., Ltd., "Face++ platform." <https://www.faceplusplus.com/face-detection/>, 2021. Accessed: 2021-07-22.
- [31] G. d. Q. J. B. S. Alvarez, P., "Riada: A machine-learning based infrastructure for recognising the emotions of spotify songs.," *International Journal of Interactive Multimedia and Artificial Intelligence. IN PRESS*, 2022.
- [32] R. Flesch, "A new readability yardstick.," *Journal of applied psychology*, vol. 32, no. 3, p. 221, 1948.
- [33] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [34] W. C. W. T. I. W. K. P. C. Y. H. . H. K. S. Chen, S. H., "Modified yolov4-densenet algorithm for detection of ventricular septal defects in ultrasound images.," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 7, pp. 101–108, 2022.
- [35] S. Hochreiter, J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [36] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, ch. 6, pp. 180–184. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [37] I. T. Jolliffe, J. Cadima, "Principal component analysis: a review and recent developments," vol. 374, p. 20150202, Apr. 2016, doi: [10.1098/rsta.2015.0202](https://doi.org/10.1098/rsta.2015.0202).
- [38] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559–572, 1901.
- [39] Instagram, "Vigo, Spain on Instagram • photos and videos." <https://www.instagram.com/explore/locations/23436873/vigo-spain/>. Accessed: 2021-07-25.
- [40] S. Bansal, C. Aggarwal, "textstat | pypi." <https://pypi.org/project/textstat/>, 2021. Accessed: 2021-07-29.
- [41] Microsoft Corporation, "Computer vision | Microsoft Azure." <https://azure.microsoft.com/es-es/services/cognitive-services/computer-vision/#overview>, 2021. Accessed: 2021-07-26.
- [42] C. Szegedy, A. Toshev, D. Erhan, "Deep neural networks for object detection," 2013.

- [43] T. Mikolov, K. Chen, G. Corrado, J. Dean, “Efficient estimation of word representations in vector space,” 2013.
- [44] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, “Distributed representations of words and phrases and their compositionality,” 2013.
- [45] L. McInnes, J. Healy, J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” 2020.
- [46] L. Van der Maaten, G. Hinton, “Visualizing data using t-sne.,” *Journal of machine learning research*, vol. 9, no. 11, 2008.
- [47] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.



Pilar Muñoz Dueñas

Pilar Muñoz Dueñas is a tenured Professor at the University of Vigo. She holds a doctorate in Financial Economics and Accounting from the University of Vigo. She has taught both undergraduate and postgraduate courses. She has written several publications and articles. She is the main researcher of the Interreg Atlantic CultureSpace project and she participates as a researcher in others at international (NTERREG POCTEPT) and national level (Retos of the Spanish Ministry of Science and Innovation).



Eugenio Doñaque Gonzalez

Eugenio Doñaque Gonzalez is a software developer and a student of informatics engineering at the Open University of Catalonia. He is currently working in the financial technology sector, and has experience with data processing systems and applied data mining and machine learning.



Ana Larrañaga Janeiro

Ana Larrañaga Janeiro is a graduate in Energy Engineering from the Universidade de Vigo (2019). She received an Interuniversity MsC in Industrial Mathematics (M2i) from the Universidade de Vigo, Santiago, Coruña, Politécnica de Madrid and Carlos III (2021), and is currently a predoctoral researcher at the Center for Research in Technologies, Energy and Industrial Processes of the same university (CINTECX), where she focuses her scientific research in the areas of Artificial Intelligence and Computational Fluid Dynamics (CFD). She focuses her research on the application of Machine Learning techniques to improve Computational Fluid Dynamics (CFD) calculations, the subject of her doctoral thesis in the Interuniversity Doctoral Program in Energy Efficiency and Sustainability in Engineering and Architecture at the University of Vigo.



Javier Martínez Torres

Javier Martínez Torres is a Mathematician and Engineering PhD from the University of Vigo. He is currently an Assistant Professor at the University of Vigo and has participated in more than 20 research projects as principal investigator. He has published more than 60 papers in JCR indexed journals and participate in more than 35 international conferences.



Ana M. Mejías

Ana M. Mejías received her PhD in Industrial Engineering from The Universidad Politécnica de Madrid (UPM-Spain). She is an Associate Professor at the University of Vigo (Spain) and she is Vice Dean in the School of Industrial Engineering. She has participated in more than 10 research projects and she has a patent in exploitation; She has published 30 papers in indexed journals and participate in more than 50 international conferences.

A Survey on Data-Driven Evaluation of Competencies and Capabilities Across Multimedia Environments

Sofia Strukova*, José A. Ruipérez-Valiente, Félix Gómez Mármol

Department of Information and Communication Engineering, University of Murcia, Murcia (Spain)

Received 22 July 2021 | Accepted 7 February 2022 | Published 3 October 2022



ABSTRACT

The rapid evolution of technology directly impacts the skills and jobs needed in the next decade. Users can, intentionally or unintentionally, develop different skills by creating, interacting with, and consuming the content from online environments and portals where informal learning can emerge. These environments generate large amounts of data; therefore, big data can have a significant impact on education. Moreover, the educational landscape has been shifting from a focus on contents to a focus on competencies and capabilities that will prepare our society for an unknown future during the 21st century. Therefore, the main goal of this literature survey is to examine diverse technology-mediated environments that can generate rich data sets through the users' interaction and where data can be used to explicitly or implicitly perform a data-driven evaluation of different competencies and capabilities. We thoroughly and comprehensively surveyed the state of the art to identify and analyse digital environments, the data they are producing and the capabilities they can measure and/or develop. Our survey revealed four key multimedia environments that include sites for *content sharing & consumption*, *video games*, *online learning* and *social networks* that fulfilled our goal. Moreover, different methods were used to measure a large array of diverse capabilities such as *expertise*, *language proficiency* and *soft skills*. Our results prove the potential of the data from diverse digital environments to support the development of lifelong and lifewide 21st-century capabilities for the future society.

KEYWORDS

Artificial Intelligence, Competencies, Computational Social Science, Data Mining, Multimedia Environments.

DOI: 10.9781/ijimai.2022.10.004

I. INTRODUCTION

HISTORICALLY, educational research has been one of the most widely explored areas to improve teaching, learning and assessment [1]. Within this context, formal education has been understood as learning that happens within regular classrooms. However, a large body of research has explored the more informal learning constantly taking place across everyday activities that fall outside of a curriculum. Furthermore, the advent of the World Wide Web, followed by the spread of Internet usage, has changed how informal learning emerges in depth [2]. In this way, it also has led to a huge growth of online learning including not only Massive Open Online Courses (MOOCs) like Coursera¹ or edX² and language learning sites like Duolingo³ or Babbel⁴, but also online portals where informal learning can emerge, such as websites that follow a Question-and-Answer format (Q&A), numerous online games, photo and video sharing platforms or social networking sites, amongst many others [3]. This kind of the platform can attract a diverse spectrum of users, especially those that

are typically less interested in traditional learning approaches [4]. By creating, interacting with and consuming the content from these environments, users can, intentionally or unintentionally, develop different skills.

All the previously mentioned environments on the web, in turn, generate large amounts of data stemming from the interactions carried out by their users in such contexts. Consequently, it is important to highlight that big data can have a significant impact on education since it offers unprecedented opportunities to support learners and advance research in the learning sciences [5]. In addition, big data practices can help discover new and useful insights about the service of education providers, its students, competitors in private sectors and, most importantly, to gain its value for better educational outcomes [6]. According to Kizilcec et al. [7], research with heterogeneous samples of learners can provide a more inclusive science of learning that moves beyond tailoring to averages. This could help us to understand a better range of issues at the core of learning and technology research [8]. These topics are connected with the term *datafication*, which allows analyses of information across large data sets in more sophisticated ways [9]. According to Mayer-Schoenberger et al., datafication refers to the transformation of social actions into quantified data that permits real-time tracking and predictive analysis [10]. This is considered as a revolutionary research opportunity to investigate human behaviour [11] and an innovative way to inspect the behaviour of learners. Hence, within the context that we are exploring, the data generated by users in these environments hold the potential to infer valuable information about users' competencies and capabilities.

¹ <https://coursera.org/>

² <https://edx.org/>

³ <https://duolingo.com/>

⁴ <http://babel.com/>

* Corresponding author.

E-mail address: strukovas@um.es

From an educational point of view, the landscape has been shifting from a focus on contents to a focus on competencies and capabilities [12], [13]. The rationale is that rapid technological evolution is having a direct impact on the skills and jobs needed in the next decade [14], [15]. Therefore, it is vital to raise a new generation that can self-regulate flexibly and rapidly acquire new skills and knowledge, as the world is changing in terms of economics, technologies, social and cultural life [16]. Accordingly, we need to focus on enhancing competencies and capabilities that will prepare our society for an unknown future during the 21st century. In this respect, interest in deploying modern techniques focused on both formal and informal learning has been rapidly growing in recent years. This has led to the fact that understanding learners and their contexts has become one of the most promising educational research topics during the past decade. For example, Redecker and colleagues [17] hypothesised that, by 2025, schools will have started integrating external learning resources and practical learning opportunities to address and implement students' individual needs and preferences. Therefore, one of the most exciting goals that researchers set in this field is to evaluate competencies and capabilities that the user potentially acquires by interacting with specific digital environments. This evaluation can be used to provide personalised feedback and to understand better how these skills develop through the interaction with these environments [18]. However, the challenge is how to perform such evaluations since there is no systematic way of doing it. Moreover, all previous studies have focused on developing and measuring competencies based on one specific platform while no one, to the best of our knowledge, studied the complete picture of the multiple existing online portals that can be used for this purpose. Hence, our survey aims to broaden the current knowledge on this issue by reviewing the research body that has already performed data-driven evaluations of competencies across different multimedia environments. As a result, this work could be the basis for developing a framework that can generalise well to different digital platforms and capabilities, and this can help to assess the capabilities of the users and to bring about a change in the educational system and training environments. Accordingly, we will try to answer the following overarching research question (RQ): Is there evidence of the existence of diverse technology-mediated environments whose data can be used to evaluate competencies and capabilities that the users can gain? For this purpose, we formulated the following specific RQs:

RQ1. What types of multimedia environments generate rich data sets that can be used for capabilities measurement?

RQ2. How are the data for the analysis accessed?

RQ3. What types of data are found across the environments?

RQ4. What methods or techniques are applied to infer competencies and capabilities?

RQ5. What competencies and capabilities are measured and/or developed across the environments?

RQ6. Are the findings across the publications validated?

RQ7. What are the main limitations or challenges faced by the authors of the studies?

The remainder of this paper is structured as follows. In Section II, we focus on the background of our study and related ones. Then, in Section III, we present the description of the RQs that drive our survey, followed by the detailed representation of the research methodology, including the description of the search process, inclusion and exclusion criteria, data collection and coding process. The research findings are outlined in Section IV, while in Section V we discuss our key findings, extend the article with a discussion beyond the results and present the implications of our research and the limitations of the selected

approach. Finally, our conclusions are presented and future research directions suggested in Section VI.

II. BACKGROUND

A. Competencies and Capabilities

Although both the terms 'competencies' and 'capabilities' are used to describe human abilities and are very closely related to each other [19], there are significant conceptual differences between them.

Generally speaking, the word competency can be defined as "something we already have or might be aiming to gain" [20]. However, the term's original definition may provide a deeper understanding of it. Initially, in 1982, a competency was defined by Boyatzis as an underlying characteristic of an individual that is causally related to effective or superior performance in a job [21]. Competencies are generally divided into functional competencies, used in daily activities, and integrative competencies, used to integrate and develop new components [22].

On the other hand, Lozano et al. [23] stated that competencies are externally demand-orientated as they are intended to provide the individual with the appropriate skills to solve external problems. In contrast, capabilities are not primarily externally demand-orientated, as they are guided by the exercise of individual freedom to choose and develop the desired lifestyle, and therefore the values individuals consider to be desirable and appropriate. Accordingly, the authors of [24] explained that an individual's set of competencies reflects their capability (i.e. what they can do) while a job competency may be a motive, trait, skill, aspect of one's self-image or social role or a body of knowledge. Consequently, capability approaches focus on developing the potential to achieve or acquire competencies, even if they are not present at a particular point in time, through certain personal qualities and attributes of individuals as well as ambition and effort [25]. At the same time, Nagarajan et al. [19] stated that capability integrates knowledge skills and personal qualities used effectively and appropriately in response to varied, familiar and unfamiliar circumstances. Finally, another important and related term is capacity. Morgan [26] stated that capacity is an emergent combination of attributes that enables a human system to create developmental value.

To sum up, competence is the quality or state of being functionally adequate or having sufficient knowledge, strength and skill, while capability is a feature, faculty or process that can be developed or improved [27]. Therefore, we can conclude that capabilities are made up of competencies that go beyond existing knowledge and experience. In this way, in our work, we refer to both, competencies and capabilities, when we use the term "capabilities". Conversely, by using the term "capacity", we refer to the overall ability of a person to achieve a goal or complete a task.

B. Related Work

We did not find any survey or literature review aiming to analyse data-driven evaluations of competencies and capabilities across several contexts at the same time. However, we did find studies that tackled one specific context or concept separately.

We found some surveys focused on detecting users online that have a high competency in specific skills, which is known in the literature as expert finding. Across these surveys, we found that one of the most established objectives within this context is detecting potential experts online. The authors of [28] aimed to review the existing literature to identify all the significant studies that addressed the task of expert finding in online communities and corporations. Accordingly to their goal, they summarised the existing graph-based and machine learning-based expert finding methods used. Similarly, Husain et al.

[29] conducted a systematic, state-of-the-art literature review of 96 articles on expert finding systems and expertise seeking. This study aimed to explore the domains that use the expertise retrieval systems, the expertise sources, the methods and the data sets used for expert finding systems. However, the most representative study related to expert finding discussed here is [30]. This comprehensive state-of-the-art survey reviewed 265 articles and proposed a framework that defines the descriptive attributes of Community Question Answering (CQA) approaches. The authors also introduced a classification of all approaches concerning problems they aimed to solve. From these studies, we can see that expert-finding is a common practice across those portals that could show users' competencies in particular topics.

Furthermore, there is a considerable amount of literature exploring the benefits of video games for developing and measuring competencies and capabilities. For example, Connolly et al. [31] examined 129 papers on computer and serious games. The game of every article was correlated with its genre, primary purpose as well as learning and behavioural outcomes. In their review, the authors found positive impacts on playing digital games on users with respect to learning, skill enhancement and engagement. The next review was conducted by Boyle et al. [32], who focused on 143 studies providing higher-quality evidence of the positive outcomes of games. While the previous work provided a useful framework for organising the research in this area, this article illustrated the increasing interest in the positive impacts and outcomes of games. Similarly, the authors in [33] investigated quality empirical studies associated with the application of games-based learning in primary education, mainly focusing on study design, the game genre, delivery platform, subject and curricular areas, as well as learning outcomes and impacts. We can see that in the field of games, the research focusing on exploring their benefits in terms of developing competencies and capabilities is fairly widespread. Moreover, we found related work performing a systematic review of publications about the potential of MOOCs in higher education aiming to investigate cognitive and behavioural learning outcomes, including obtained or mastered knowledge and intellectual skills [34].

The above review reveals that all previous studies explored how to develop and measure competencies based on one specific type of online environment. However, no single study, to the best of our knowledge, has provided a full picture of the multiple existing multimedia environments that can be used for this purpose. With this objective in mind, we surveyed the existing literature to study data-driven evaluations of competencies and capabilities across various digital environments.

III. METHODOLOGY

First of all, we stated the RQs and, accordingly, several steps were taken to select the publications for our survey, including (1) a literature search for identifying relevant articles, (2) the inclusion and exclusion criteria, (3) a full paper review and coding process, and (4) synthesis and analysis. The entire methodology process is represented in Fig. 1.

A. Search

Our literature search entailed an ill-defined area arising from many research communities. This led to several challenges regarding setting a clear search methodology to conduct the survey. The main challenge was that we could not directly retrieve all the literature on this topic by simply using a set of keywords and searches. Therefore, we were not able to apply an utterly systematic search as proposed in the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement [35]. On the other hand, while conducting our non-systematic literature review, we mainly relied on a knowledgeable selection of relevant current, high-quality articles [36]. Although we had to adapt the search process to the context of the target research

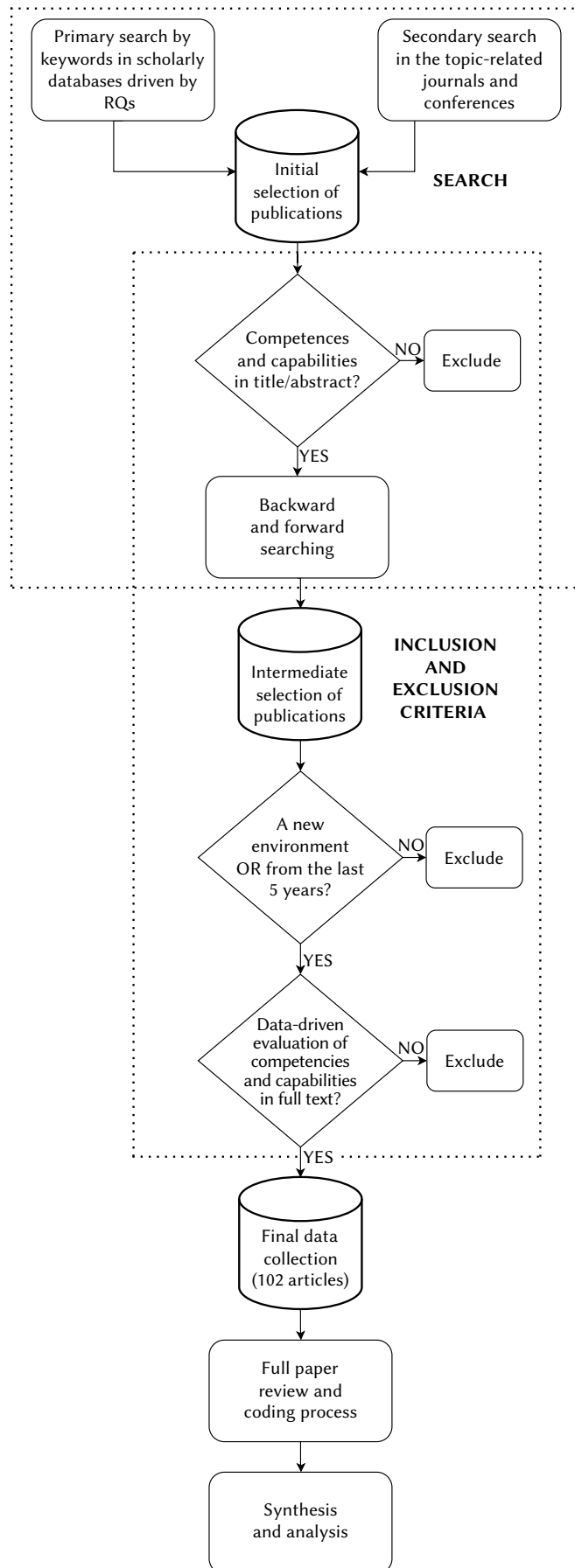


Fig. 1. Overview of the methodology used to conduct the survey.

area and, accordingly, change it, the rest of the methodology followed the PRISMA guidance and exemplars.

Initially, we identified keywords and formulated search strings according to our goal of exploring diverse multimedia environments in which data can be used to perform a data-driven evaluation of competencies and capabilities of different nature. We started the literature search with the following approach:

1. We looked for related papers using keyword search on various indexing platforms such as Scopus⁵ and Google Scholar⁶. The examples of the keywords are as follows: “expert identification,” “information retrieval,” “game-based assessment,” “forum,” “question and answer portal,” “online learning,” “MOOC,” “competence,” “capability,” “skill,” “engagement,” “soft skills,” “language proficiency,” “correct on/at (the) first attempt”.
2. We explored the publications of the corresponding top conferences and journals related in the fields.

Firstly, we performed a fast initial paper screening to discard those that did not fit the survey by reading their title and abstract. More specifically, a paper was not included if there was no mention of the potential measurement and/or development of competencies and capabilities in its title. This probably meant that its primary focus was unrelated to our interest. In case the title, together with the abstract, did not provide sufficient information to justify its exclusion, the paper was not excluded. Next, we conducted the so-called ‘snowballing technique’ [37] comprising two methods: backward searching and forward searching. Backward searching included the review of the references of the publications fitting our survey while forward searching consisted in the examination of the papers that cited a known and relevant publication. We applied these methods to the publications that already proved to belong within our survey. Finally, the newly selected papers went through the same screening criteria as initially described by reading their title and abstract.

B. Inclusion and Exclusion Criteria

After performing the search and identifying the papers fitting our survey, we formulated several mandatory criteria for the final decision of whether to include the paper in the survey or not. Creating a valid set of inclusion and exclusion criteria required considerable trial and error pilot testing [38]. Accordingly, a paper was excluded if none of the following conditions was met (in other words, if at least one condition was met, the paper was included):

- The paper was published in the last five years.
- The paper represented a new environment or features that would deepen the results of the RQs.
- The paper was peer-reviewed.
- Reported outcomes in the paper included all the data needed for answering RQs outlined in Section I and were presented appropriately and consistently.
- The data from one paper did not overlap with the data from another paper.

⁵ <https://scopus.com/>

⁶ <http://scholar.google.com/>

We used, as a quality proxy, the type of publication and the venue, including full papers in conferences, book chapters and articles published in reputed journals. After that, we performed a more careful review of those papers originally selected. The abstracts were read again together with the methods and results sections. Through this process, we ensured that the selected publications measured concrete competencies and capabilities and performed their data-driven evaluation and thus correctly fitted the survey.

C. Data Collection

The final data collection included a total of 102 articles out of hundreds of initially selected publications. These articles are presented in ascending chronological order in Appendix B.

Most of the publications have a selection of keywords that define the research topic. We collected the keywords of all the papers selected for the survey, and we counted how many publications used the same keywords. The most frequent keywords are presented in Fig. 2.

The total sum of keywords is 415, while there are 329 different keywords. The average keyword was found 1.6 times, and its variance is 1.45. Based on the keywords (see Fig. 2) which define the publications’ main topics, we did observe a high variability in the type of keywords used. Therefore, we see that the research topic covers many areas.

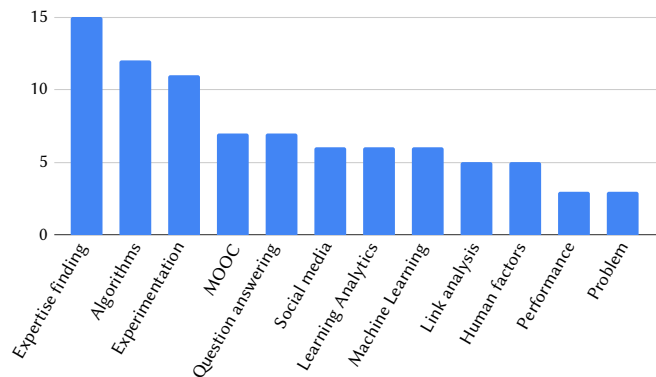


Fig. 2. Distribution of keywords across the articles selected for our survey.

D. Full Paper Review and Coding Process

In the coding stage, we identified variables that we consider as the most valuable to address the RQs stated in Section I. Based on the aim of our survey and the fact that we did not know in advance what data we could find, we followed an inductive coding scheme. This means that codes were created based on the qualitative data themselves. In Table I, we outline the variable coding scheme that we followed in our survey, indicating each code with its possible labels, which represent qualitative data. In the event of a paper deploying several experiments or using multiple methods [39]–[41], these were coded with several variables such as “*Network analysis*, *NLP*”. It is also important to mention that in the case of multiple methods, we coded only those used to infer users’ competencies or capabilities rather than to report results. The full results of the coding process per paper are presented in Appendix B.

TABLE I. THE VARIABLE CODING SCHEME

Environment (RQ1)	Data access (RQ2)	Data type (RQ3)	Methods (RQ4)	Capabilities (RQ5)	Validation (RQ6)
Content sharing & consumption	Open data set	Textual	Machine Learning	Expertise	Yes
Games	Public domain	Biometric	Statistics	Language proficiency	No
Online Learning	API	Clickstream	Network Analysis	Soft skills	
Social Network	Direct access	Audiovisual	Experiment/control	Performance	
			Natural Language Processing	Cognitive skills	

At the same time, from the coding process, three groups of publications emerged depending on their degree of relationship with the main goal of our survey:

1. The **strong relationship** group contains those papers that exactly matched the topic of the survey. This means that they proposed a method for identifying competencies or capabilities based on the data gathered by the user's interaction with a particular digital environment. We assigned 56 papers to this group.
2. The **weak relationship** group is similar to the first one; however, instead of measuring capabilities, the work belonging to this group explored other data-driven behavioural characteristics of users related to capabilities such as engagement or influence, to name some examples. Although this specific behaviour was not within the primary focus of our survey, these publications might be useful for analysing the applied methods and the selected environments. For example, several studies have explored a user's influence on others [42]–[46] or predicted dropout rate [47], [48]. We assigned 31 papers to this group.
3. The **high potential** group comprises those papers that had goals less related to our survey objectives but still described thought-provoking multimedia environments with large amounts of user data that could be used to perform a data-driven evaluation of capabilities. This group includes mainly those studies which did not provide evidence of using data generated through the user's interaction with the environment; instead, these studies measured other characteristics of users not related to their competencies or capabilities (identified in RQ5). We discuss these studies in more detail in Section B. We assigned 15 papers to this group.

IV. RESULTS OF THE SYNTHESIS AND ANALYSIS

In this section, we highlight our findings discussing each RQ in detail. We start with a description of all the types of environments within the scope of our survey. Next, we describe the data access, followed by the data type and the applied methods that emerged after analysing the articles. Finally, we discuss particular competencies and capabilities measured and/or developed across the selected environments. We also present the analysis of the validation of the results across the selected studies, followed by the main limitations mentioned by the authors of the publications.

A. What Types of Multimedia Environments Generate Rich Data Sets That can be Used for Capabilities Measurement? (RQ1)

Four groups of environments were identified during the review process described in the methodology section, namely, **content sharing & consumption** (41 articles), **video games** (13 articles), **online learning** (31 articles) and **social networks** (16 articles). They can all generate rich data sets through the users' interaction, which can be used to explicitly or implicitly perform a data-driven evaluation of competencies and capabilities. Next, we describe in more detail each one of them.

1. Content Sharing & Consumption

In this group of environments, we have all the platforms where users either intend to share content or consume it. Their categorisation is as follows:

- **Q&A portals and forums.** They include Quora⁷ [49], Reddit⁸ [50], StackOverflow⁹ [51], [52], and Yahoo! Answers¹⁰ [53], among

others. These environments are quickly becoming rich sources of knowledge on many topics not well served by general web search engines [54]. They can be defined as online discussion sites where users can post messages stating what they are interested in or replying to others. On these platforms, users can ask questions, give answers and also provide their assessments about the quality of questions or answers through votes and choosing favourites [55].

- **Photo and video sharing online platforms.** Hosting services for video are called Online Video Platforms (OVPs); they allow users to upload, play, store and transfer video content online. These include, among others, YouTube¹¹ [56] and Vimeo¹². On the other hand, users of Instagram¹³ [57] can share both photos and videos, and Pinterest¹⁴ allows to share only images so that users create, organise and share content by creating visual bookmarks called pins [58], so it is possible to characterise the volume and coherence of users' pinning activity in a given category [59].
- **GitHub¹⁵.** This platform [60] does not match any category above, but it still represents an appropriate and important portal corresponding to the stated goals. This service allows its users to host open-source projects and work collaboratively on them.

It is also worth mentioning that some papers discussed several *content sharing & consumption* platforms in their research scope. For example, the authors of [61] presented a method to create a detailed technology skill profile of a candidate. This was based on his code repository contributions through annotating user contributions on GitHub code repositories with technology tags found on StackOverflow.

2. Video Games

Even though the majority of *video games* are set out to entertain, nowadays there are many games created for other purposes, and their value as a learning tool has been widely accepted [62]. The explored *video games* that we found are categorised according to their main purpose into:

- **Entertainment games.** These are the commercial games that are played for entertainment purposes. For example, [63] described a racing game and explored the creation of its players' engagement profiles. The games of this type have an easily understandable user interface, their main goal is to entertain, and they are designed to provide an immersive experience for the player through gameplay, narrative and challenges [64]. Moreover, this kind of game can help develop various competencies and capabilities (e.g., multitasking and problem-solving skills) without users even noticing their involvement in the educational process.
- **Educational games.** This type of game is designed for educational purposes, and numerous teachers have already tried to teach foreign languages, history or programming with these games [65]. The use of educational games attempting to digitise and optimise the learning process has increased significantly [66]. This is due to the initial investigations suggesting that it can be rewarding at many levels, including academic achievement, concentration and classroom dynamics in general. Many experiments were conducted seeking to create *video games*, which could be beneficial for the learning process [67]. After the first studies succeeded, an interest in educational *video games* attracted the attention of many researchers. Nowadays, it is an up-and-coming field opening new ways of conducting educational processes [68].

¹¹ <https://youtube.com/>

¹² <https://vimeo.com/>

¹³ <https://instagram.com/>

¹⁴ <https://pinterest.com/>

¹⁵ <https://github.com/>

⁷ <https://quora.com/>

⁸ <https://reddit.com/>

⁹ <https://stackoverflow.com/>

¹⁰ <https://answers.yahoo.com/>

According to the stated RQs, one example we were interested in is [69]. This article described the Virtual Age game, which is effective for learning about evolution, as its authors claimed. For this purpose, the scientific concept of evolution was implemented, concretised and gamified. The game's main idea is to create some extinct creatures, keep them alive and transmit them from one era to another. The results proved that Virtual Age is useful for learning about biological evolution.

- **Serious games.** These games are designed for purposes other than, or in addition to, pure entertainment. In other words, they can train specific skills and improve learning performance through real-world problem-solving [70]. The authors of [71] consider serious games as adaptive systems, as they continually adjust their responses to the learners' actions for preserving favourable conditions for playing and learning. The difference with educational games is that serious games are not directly connected with educational goals but can aim to change the users' behaviour, attitude, health or other features. Still, the educational purpose is not excluded.

In this paper, we consider serious games as a subtype of educational ones. For example, a serious game described by Kang et al. [70] provides an authentic learning context for space science and astronomy. Users have to find a suitable home in the solar system to relocate a group of six distressed aliens because their home planets have been destroyed. The authors provided an analytical approach to understanding in depth students' sequential patterns of behaviour. At the same time, they showed that problem-solving strategies were different between low- and high-performing students.

3. Online Learning

Several types of online learning environments emerged after analysing the surveyed articles:

- **MOOCs (Massive Open Online Courses).** They are defined by Daradoumis et al. [72] as one of the most versatile ways to offer access to quality education, especially for those residing in far or disadvantaged areas. Taken as a whole, MOOCs can provide not only traditional learning materials but also make students' forums or social media discussions available. There are many studies examining data from various MOOCs although, in our work, we mostly focus on the most well-known MOOC providers such as Coursera [73]–[76] and edX [77]. These MOOCs follow a similar way of providing the content of the courses, evaluating the received knowledge and, most importantly, sharing the same goal.
- **Language learning** websites and applications. Their main goal is to improve the language proficiency of their users. For example, one of the portals matching the criteria is Duolingo, a popular language-learning platform that applies gamification techniques [78].
- **Traditional formal online learning.** Here we include learning management systems, educational software platforms as well as virtual learning environments whose main goal is to support teaching and learning. There are several articles discussing the benefits of applying these systems such as Moodle¹⁶ [79] and WebCT¹⁷ [80].
- **New applications.** In this group, we classify other online portals that do not fit in any of the previous categories. Here we include all *new applications* developed for learning [81] which can be classified neither as MOOCs nor as language learning or traditional

learning. Moreover, in this group, some tools facilitate online learning. For instance, the authors of [82] presented an intelligent mobile learning system optimised for consuming lecture videos in both MOOCs and flipped classrooms. Another interesting example is an E-book system that allows students to preview their lessons before the class and to write questions. Moreover, they can take notes, mark part of a page as important content during the class and review the learning content after classes [83]. On the other hand, during the class, a teacher can monitor how many students are viewing the same page as the teacher, whether they are following the explanation or whether they are reading previous or subsequent pages [84].

4. Social Networks

Social networks are defined by Boyd et al. [85] as web-based services that allow individuals to (1) construct a public or semi-public profile; (2) articulate a list of other users with whom they share a connection; and (3) view and traverse their list of connections. These services have proved to be an exceptionally useful tool to interact with other people [86]. This has led many researchers to use them to find various trends, among which we can highlight Facebook¹⁸ [87], [88], Twitter¹⁹ [41] and LinkedIn²⁰ [89].

By way of example, the authors of [90] worked on understanding the use of *social networks* and *video games* among teenagers. They concluded that the respondents felt the desire to satisfy their needs for pleasure and communication by using these environments, without assuming they were willing to merge it with the educational process. However, several studies proved that this is not an unachievable goal. For example, Boukes [91] investigated how the use of Twitter and Facebook affected citizens' knowledge acquisition. In addition, there is research [92] done on analysing the effect of *social networks* engagement on cognitive and social skills.

5. Summary

Fig. 3 presents the distribution of the environments across the selected articles. The *content sharing & consumption* environment prevails in most published studies since there is a significant amount of research done on this topic covering various platforms.

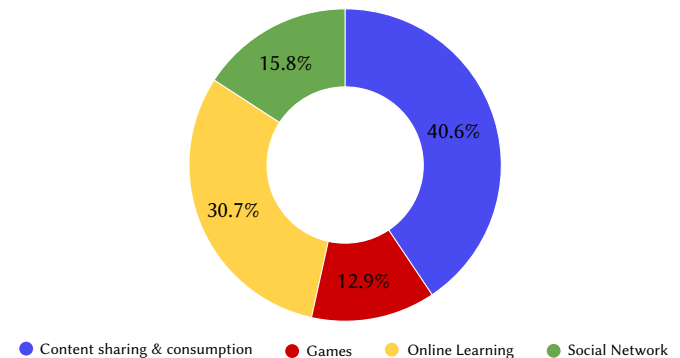


Fig. 3. Environments measuring and/or developing capabilities.

B. How are the Data for the Analysis Accessed? (RQ2)

The action 'to access data' is generally understood as the ability to retrieve, modify, copy or move data from different sources. The analysis revealed that there are four main ways of accessing data in the publications that we selected:

¹⁸ <https://facebook.com/>

¹⁹ <https://twitter.com/>

²⁰ <https://linkedin.com/>

¹⁶ <https://moodle.org/>

¹⁷ <https://blackboard.com/>

1. **Direct access.** Having *direct access* to data means that the researchers own or were provided with the data. For example, this can happen if the researcher is working for the company [93], has direct access to the systems' database [94] or is asking the organisation that owns the data for permission to access them [69].
2. **Application Programming Interface (API).** An *API* is a set of functions that allows building and integrating the software of applications. There is a variety of public APIs that we can interact with. For instance, the Twitter API21 enables programmatic access to Twitter in unique and advanced ways that can be used to analyse, learn from and interact with tweets, direct messages, users and other key Twitter resources. For example, data were accessed through an *API* by Bouguessa et al. [95] and Bigonha et al. [44]. Finally, some studies have focused on other APIs such as the Graph API of Facebook [96], Reddit API [50], Yahoo! Answers API [53] or GitHub REST API [97], amongst others.
3. **Public domain data.** Data are available in the *public domain* where researchers can easily access them. For example, the authors of [98] collected hundreds of thousands of weblog files from Coursera using data mining techniques. Similarly, Han et al. [99] developed their own web crawling software since Pinterest does not provide an official API for data collection, and Calefato et al. [100] developed a custom web scraper for Apache Software Foundation. However, it is worth mentioning that scraping large-scale social media data from the web requires a high degree of engineering skills and computational resources [101].
4. **Open data set.** Data published as an *open data set* means that anyone can freely access these data. The difference between *public domain* data and *open data set* is that the latter is structured, well-maintained data, and released under certain public licenses that specify how others can use it. Moreover, *open data sets* are easier to access; they are already clean and ready for analysis. For this reason, there is a significant amount of papers [40], [102], [103], which simply downloaded *open data sets* for conducting their research.

Fig. 4 summarises the results of the data access across the papers. As hypothesised, our findings show that a significant number of publications had *direct access* to data or used *public domain* data. Most papers either decided to use the generated data to objectively analyse the outcome and the benefits of the product or chose to use already available data in the *public domain*.

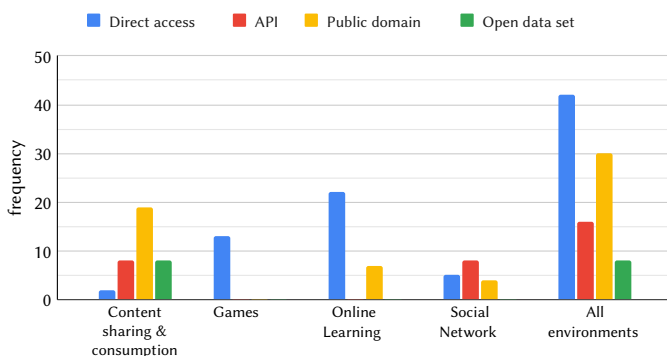


Fig. 4. Data access distribution across the selected articles.

C. What Types of Data are Found Across the Environments? (RQ3)

Initially, we aimed to examine diverse technology-mediated environments that can yield rich data sets generated by users in these environments. Therefore, data play a significant role in our survey

because these data hold the potential to infer valuable information about users' competencies and capabilities. Given the aforementioned methods to access data, the developed capabilities of users were retrieved by analysing various types of data, including:

1. **Textual data type.** It is the most common type of data used among the selected publications. Most of the work (e.g., [43], [44], [60], [104]–[109]) mentioning its use has in common that it employed *textual* data generated directly by users – questions and answers, posts, tweets or source code. At the same time, other work [110], [111] generated or extracted *textual* data describing the users' statistics, including active time, the number of completed lessons, amount of lessons per day, test scores, test solutions, successes, difficulties, course progress, etc.
2. **Clickstream data type.** It represents the mouse clicks made by the user when interacting with the web environment. This is a widespread data type observed across environments – there are several studies in games, e.g. [63], [65], [112], using students' *clickstream* log files to explore their behaviour. Another example is [66], where authors extracted different behavioural features from students' evidence trace files logged by the game server such as their actions, time and performance on each specific assessment task, as well as general behavioural features not specific to assessment tasks and students' pre-test scores. Analysing this kind of data could be useful for improving online education for both teachers and students [113] by understanding how students use course resources, what contributed to their persistence and what advanced or hindered their achievements [114]. Another powerful use of *clickstream* data is described by the authors of [115], who detected cheating in MOOCs using a rule model based on heuristics.
3. **Audiovisual data type.** These data include videos, images and audio files. These data can be retrieved from any type of environment and, with the right methods, they can be informative. For example, the study described in [116] was carried out using *audiovisual* data for automated prediction and analysis of job interview performance while the authors of [81] created a system called Social Skills Trainer. It consists of a virtual avatar that recognises the user's speech and language information and can provide feedback to users to improve their storytelling skills.
4. **Biometric data type.** It is a body characteristic that can be measured or calculated. Theoretically, these data can be retrieved across different contexts; however, sophisticated instrumentation is needed to measure these data. For example, it is common to use sensors to detect eye-gaze or wearables to measure heartbeat and electrodermal activity (EDA). Therefore, researchers need to aim specifically to acquire such data as part of their research process since these data cannot be accessed easily. We found evidence of *biometric* data only in *online learning* environments for research purposes. In our context, biometric data could be useful for a better understanding of student engagement or stress resistance.

The most representative example of the research evaluating *biometric* data is the work carried out by Peng et al. [117]. The authors proposed that, by monitoring Heart Rate Variability (HRV), it is possible to detect a person's cognitive performance. The experiment consisted in measuring the HRV of participants while they were taking part in the discussions and, through this kind of data, evaluating the discussion skills. Next, several models were adopted, such as Logistic Regression, Support Vector Machine and Random Forest. The results proved that participants' HRV data could effectively evaluate the answer quality of Q&A segments as an automatic evaluation of participants' discussion performance. This method outperformed compared with using traditional Natural Language Processing such as semantic analysis.

²¹ <https://developer.twitter.com/en/support/twitter-api>

A visualisation of data types per environment is illustrated in Fig. 5. We can note that the vast majority of publications in environments such as *content sharing & consumption* and *social networks* use textual data while *video games* and *online learning* environments primarily generate *clickstream* data. However, all the environments mentioned previously might contain *audiovisual* data including images, videos, audios and *biometric* data such as heart rate [82] or fingerprints. Also, in general, we see a clear trend of the superiority of the *textual* data across most environments. At the same time, it is curious to mention that *online learning* environments can produce all types of data. This is probably because, for this group, we refer not only to traditional formal online learning and MOOCs but also to language learning platforms and all applications aiming to educate their users. These categories are very innovative; therefore, we could intend to use all possible data types.

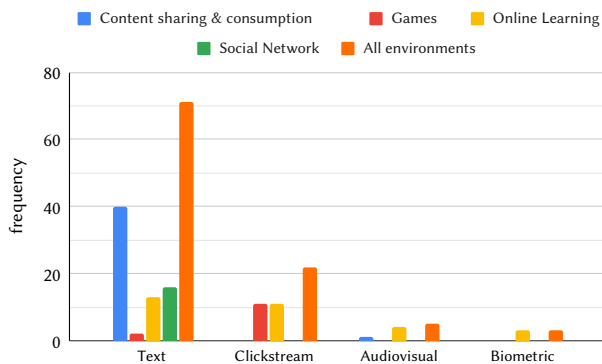


Fig. 5. Data types per environment.

D. What Methods or Techniques are Applied to Infer Competencies and Capabilities? (RQ4)

After exploring what types of data were retrieved across the publications, our goal was to examine the methods that were applied for their analysis. Accordingly, we found that several groups of methods used for analysing the data emerged. They include:

1. **Statistics.** They encompass different mathematical analyses covering many methods, tests and metrics, including Poisson models [118], Mann-Whitney U tests [119], Analysis of Variance (ANOVA) [70], Spearman's Rank Correlation Coefficient [46], Student's t-tests [102], among many others.
2. **Machine Learning (ML).** It is an Artificial Intelligence application by means of which systems can automatically learn and improve from experience without being explicitly programmed for a specific goal. Some of the most common ML algorithms include NN, Support Vector Machine (SVM) [120], Naïve Bayes, Logistic Regression [89] or Random Forest [97]. Many of the analysed publications, e.g. [82], [105], [117], apply a number of these algorithms to compare their performance.

The authors of [100] performed a quantitative analysis of data from the code commits and email messages contributed by the developers working on the large-scale distributed projects of Apache Software Foundation. They aimed to find evidence that personalities can explain developers' behaviour. The authors applied a Principal Component Analysis (PCA) to group developers with similar personalities, and cluster analysis was performed to reveal developers resembling each other but also differing from the rest. For building the contribution likelihood model, a logistic regression model was used. One of the conclusions was, for example, that the propensity to trust others turned out to be positively influential on the result of code reviews in distributed projects.

3. **Network analysis.** It is an analytical method used to evaluate relationships between nodes that are a part of a connected network. Bouguessa et al. [121] stated that most of the existing approaches attempting to discover experts model the environment as a graph in which the nodes represent users and the edges represent the interactions between them. In this way, the authority score is generally measured through graph-based ranking algorithms such as PageRank, Hyperlink-Induced Topic Search (HITS), InDegree Algorithm, etc. [122].

In turn, the authors of [123] described the idea underlying the HITS algorithm as follows: "A good authority is one that is pointed to by many good hubs, and a good hub is one that points to many good authorities." [p. 238] In simple words, the quality of a page as an authority depends on the quality of the pages that point to it as hubs and vice versa. The HITS algorithm was used by Jurczyk et al. [54] for predicting experts in Q&A portals, and its effectiveness was proved by performing a large-scale empirical evaluation.

4. **Natural Language Processing (NLP).** It combines various techniques from different computer science areas, linguistics, and artificial intelligence to interpret, process and analyse human language, often on a large scale. One of the papers using NLP is [50], where authors performed a semantic analysis by using a text analysis software called Linguistic Inquiry and Word Count, which counts words that belong to psychologically meaningful categories.
5. **Experimental design.** This approach can be used when researchers are interested in evaluating the impact that specific design decisions or characteristics can have on different outcomes. For example, they can conduct an experiment dividing participants into two groups, where the experimental group is the one that tests a new feature, whereas the control group does not test it. For example, Lesser [64] aimed to determine the effectiveness of digital game-based learning compared to other teaching methods related to music education. An experimental study consisting of test and control groups together with in-depth interviews led to the following results: students who had access to educational *video games* combined with the assistance of an instructor achieved higher mean scores compared to students who had access to either *video games* without instruction or instruction without *video games*. This author suggested that educational *video games* may be potentially used as an effective tool in the music classroom to teach musical concepts and skills as well as to increase student motivation, engagement and a hands-on approach to learning.

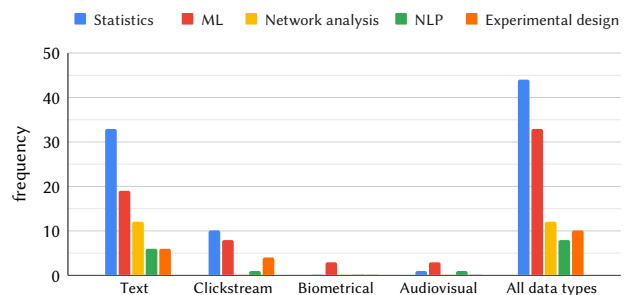


Fig. 6. Methods per data type.

A visualisation of data types per method is illustrated in Fig. 6. It shows that many publications applied statistical methods and ML models. This is because statistics and ML are broad fields that include a variety of different subfields. Next, we analysed how the applied methods spread across the data types. As we see, on *textual* data, all the identified methods were performed. This can be explained by the fact that many more publications focus on this type of data, and therefore there are many more examples of it. Besides, using *textual*

data by default implies the possibility of applying various methods. Moreover, we can notice that *ML* is applied to all types of data. With a huge development of this field and the improvements of its methods, their utilisation across various data types has become increasingly widespread.

E. What Competencies and Capabilities are Measured and/or Developed Across the Environments? (RQ5)

For this RQ, we examined diverse technology-mediated environments with the ability to generate rich data sets through the users' interaction. We observed that these data could be used to explicitly or implicitly perform a data-driven evaluation of capabilities. In this section, we summarise the main competencies and capabilities that emerged from the survey, namely, **expertise**, **language proficiency**, **soft skills** and patterns of **behaviour** that were targeted across the different studies.

1. Expertise

Expertise is defined by Herling [124] as the optimal level at which a person is able and/or expected to perform within a specialised realm of human activity. This broad group of capabilities includes:

- **Computer science expertise.** It is defined as proficiency in different programming languages, libraries or tools. The studies aiming to evaluate such skills mainly focused on portals highly related to the field of computer science such as GitHub [97], [125]–[127], StackOverflow [51], [128], [129] or both of them simultaneously [130]. In general, many researchers concluded that information technology capabilities are key drivers to achieve superior customer relations and innovation [131].
- **Topical authority** in a selected topic. The most suitable portals for finding this kind of *expertise* are forums and Q&A websites since they already focus on a particular topic and their users generate a substantial amount of data suitable for the analysis. This analysis revealed that the following portals have been used for this research objective: Yahoo! Answers [121], [132], Quora [49], [133], Reddit [103], AskMe forum²² [134], TurboTax Live Community²³ [135], MedHelp²⁴ [136] and Tianya Wenda²⁵ [137].

There are many examples of successful work using data to detect *topical authority* in a selected topic with significant results. The most representative one is the research conducted by Abdaoui et al. [138]. The authors aimed to detect posts written by medical experts in health forum discussions. They managed to collect more than 28,000 textual messages from two specialised websites. Through these data, a supervised learning approach to distinguish posts written by medical experts and by patients in health forums was followed. This research shows that it is possible to detect topical experts.

- **Learning outcomes metrics.** We can find this kind of *expertise* only in *online learning* environments since these are the only ones providing tasks to students and evaluating the results. For example, Brinton et al. [139] presented two frameworks whose purpose is to represent video-watching clickstreams: one based on the sequence of events created and another on the sequence of viewed videos. The authors extracted students' actions such as reflecting (repeatedly playing and pausing) and revising (plays and skip backs). The authors concluded that some of this behaviour is significantly associated with user performance in online learning.

²² <https://ask.metafilter.com/>

²³ <https://ttlc.intuit.com/>

²⁴ <http://medhelp.org/>

²⁵ <http://wenda.tianya.cn/>

The aforementioned portals generate a significant amount of data, which can allow the detection of potential experts. Although this large volume of user-generated content is a potential strength, it also makes the problem of finding authoritative users for a given topic challenging [57]. According to Riahi et al. [39], experts are often not provided with questions matching their expertise and, therefore, new questions may not be matched with an expert properly; and hence they end up without receiving a proper answer. For this reason, improving expertise finding algorithms would be useful to enhance the user experience in these portals.

2. Language Proficiency

Language proficiency refers to a person's ability to correctly use a certain language in terms of grammar, fluent speaking, lexical understanding, etc. We have found two main approaches to infer this capability:

- In environments specifically designed for developing language skills, i.e., language-learning portals such as Duolingo or Babbel, whose main purpose is to help users to fully learn a language through specifically tailored online activities and courses.
- When researchers collect data from other environments where users can exhibit evidence of *language proficiency* capabilities. For example, the authors in [140] investigated whether Twitter could support creative writing development and in [141] English language learners' use of Instagram was explored.

While we found multiple studies detecting improvements in learning foreign languages when learners use various *social networks*, these studies frequently used *experimental design* to measure the impact of *social networks* on the use of language learning [87], [142]. Therefore, they were not directly using the data from learners to evaluate their language capabilities, and consequently, these studies are only weakly connected to our goal.

By way of conclusion, we did not find many studies that firmly fit our survey for this particular capability. One of those that fit well, however, is a spoken *language proficiency* assessment system called Dolphin [143]. Its goal is to automatically evaluate students' phonological fluency and semantic relevance by analysing students' video clip and verbal fluency tasks. The results proved that Dolphin could provide more opportunities to practice and improve their oral language skills, and at the same time, it could reduce teachers' grading burden. The experiments demonstrated the effectiveness in model accuracy, system usage, teacher satisfaction rating and other metrics.

3. Soft Skills

Soft skills are described as a combination of interpersonal and social skills [144]. They are used to indicate personal transversal competencies and personality traits that characterise relationships between people [145]. We found several studies that can be divided into the following *soft skills* capabilities:

- **Executive control skills.** They are defined by Strobach et al. [146] as the control and management of other cognitive processes as well as cognitive skills, working memory and attention skills. One of the studies exploring the relationship between executive control skills and action *video games* is [146]. In this work, video gamers were shown to improve performance in dual-task and task switching situations in comparison with nongamers. Another sample study measuring cognitive skills, working memory and attention skills is the work carried out by Alloway et al. [92]. They measured the impact of *social networks* engagement on cognitive skills and social connectedness.
- **Creativity.** It can be understood as the ability to generate ideas that could be beneficial for problem-solving, communication and

entertainment. The authors of [147] stated that creativity could be inculcated, encouraged and trained. For this purpose, they developed a digital game-based learning system to foster students' creativity.

- **Problem-solving skills.** They represent a range of attitudes and thinking skills that are used to find solutions to problems [148]. Chu et al. [149] proposed a game-based development approach for improving these skills. They conducted an experiment in an elementary school natural science course aiming to evaluate the performance of their approach. Finally, it was proved that the proposed game development-based learning approach could effectively promote the students' problem-solving skills.
- **Critical thinking.** It is defined as 'the intellectually disciplined process of actively and skilfully conceptualising, applying, analysing, synthesising and/or evaluating information gathered from or generated by, observation, experience, reflection, reasoning or communication, as a guide to belief and action' [150]. Chootongchai et al. [79] stated that it is crucial to have a range of thinking and innovation skills, including critical thinking, collaboration and communication, to be successful at work and in life more generally. Accordingly, they developed an online learning system to enhance thinking and innovation skills for higher education learners.
- **Social skills.** We could only find one example of a study evaluating social skills. The authors of [81] developed a dialogue system called Automated Social Skills Trainer that can decrease human anxiety and discomfort in social interaction and help acquire social skills through human-computer interaction. The system includes a virtual avatar that recognises user speech as well as language information and gives feedback to users to help them improve their social skills.

4. Behaviour

Within this group, we mainly consider various behavioural patterns, including engagement, influence or dropout, amidst others. We cannot name them as capabilities, but they are essential for our survey because, as stated in Section B, through publications describing *behaviour*, we can learn of additional studies that hold the potential to perform a data-driven evaluation of competencies.

5. Summary

The distribution of the developed capabilities per environment is represented in Fig. 7.

We observe that all the stated environments prove to measure capabilities of a different nature. We also see that *content sharing & consumption* environment prevails in the development of expertise. This is because, in such environments, we consider many forums and similar websites where it is possible to observe experts from different fields.

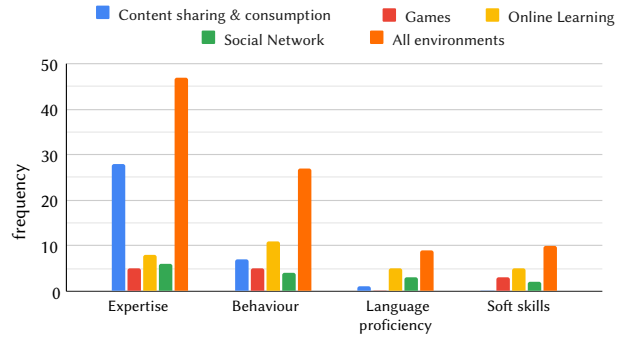


Fig. 7. Capabilities per environment.

F. Are the Findings Across the Publications Validated? (RQ6)

One of the most significant parts of the research reviewed in our survey is the validation of the results. Validation is intended to ensure that the proposed methods and their results proved satisfactory by conducting appropriate experiments. According to our survey topic, only the validation made by humans was considered since it has high reliability and validity.

A close inspection revealed that up to 81.4% of all the analysed studies do not validate their results. After a careful analysis of the publications that did not validate their results, we can conclude that the results' validation was beyond their research goals. Another possible explanation could be that it was not feasible to validate the results. For example, it is not a trivial task to verify that soft skills such as adaptability or stress management skills were obtained.

Regarding the publications that did validate their findings, we saw that almost all of them verify the expertise search results (see Fig. 8). Expertise has an understandable way of validation. The most common one is described in [151]. This paper presented an approach to identify potential experts. The most noteworthy aspect of this work is that the authors also proposed a method to detect users likely to become experts in the future through their behaviour and estimation of their motivation and ability to help others. After building the models and having the results, a human evaluation was performed. The authors asked community managers to evaluate the potential experts identified by the algorithm, and the analysis revealed that there is a high agreement between the human evaluation and the performed algorithms.

Another interesting approach to performing validation is through another portal. For example, the authors of [130] followed a two-step approach. The first step was to measure developers' commit activity on GitHub by considering both the quantity and the continuity of their contributions in isolated projects over time. The second step was to evaluate the generated developers' expertise profiles against recognised answering activity on StackOverflow via a data set of users that were active both on GitHub and on StackOverflow.

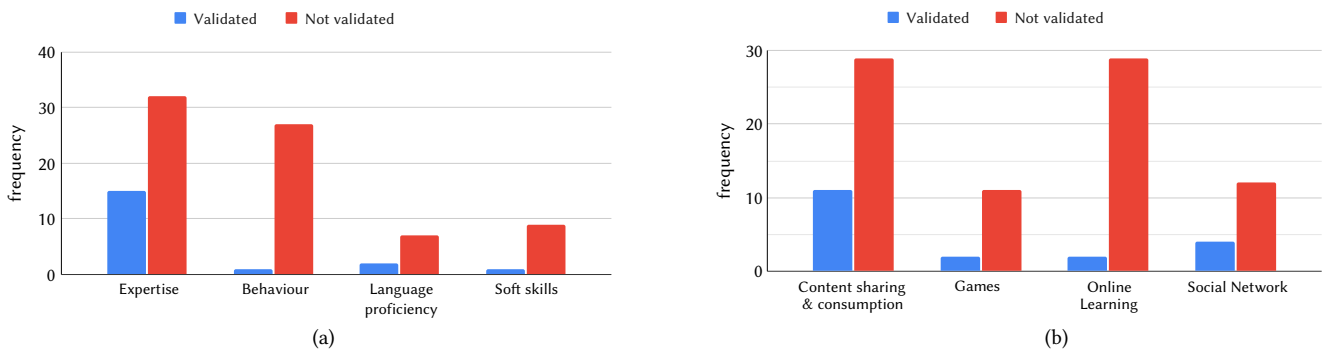


Fig. 8. Validation results: (a) per capability and (b) per environment.

To sum up, it is important to conclude that we did not find many studies performing manual validation of their results. This could be related to the fact that human validation requires much more time and effort and, therefore, the researchers decided to give preference to other methods. As we can also see in Fig. 8, validation by humans was performed mainly in *expertise finding* through the *content sharing & consumption* environment.

G. What are the Main Limitations or Challenges Faced by the Authors of the Studies? (RQ7)

Limitations show potential weak points of the study, and researchers can encounter them due to constraints in research design or methodology. We can group the limitations that the authors faced as follows:

- **Data.** Data were crucial for our survey because our goal was to explore diverse multimedia environments generating rich data sets through the users' interaction and where these data can be used to perform a data-driven evaluation of capabilities. Therefore, we examined the limitations related to data in much detail and reached several conclusions. Firstly, data access limitation is connected with difficulties encountered during the data retrieval, such as being limited in accessing data or not having the right to do so. One of the articles dealing with data access limitation is [152]. In this study, the authors claimed that many researchers who used data extracted from platforms with APIs faced the same limitation. Secondly, the authors of [151] stated that for expertise finding, they only considered a single Q&A site with a narrow purpose and an active team of professionals behind the scene. At the same time, a vast majority of studies, e.g. [44], [88], [104] aim to use more data of the already selected portal or to apply their method in a data set from a new portal as a part of their future work. One more issue of not having large enough data sets can be that results cannot be accurate and reliable enough [93]. Therefore, it is vital to have enough data and to know how to deal with inaccurate, misleading or biased data [56]. In total, 13 publications raised this limitation.
- **Methods.** This limitation was also crucial since researchers' results highly depend on the applied methods to measure the capabilities. There are several studies [93] stating that one of their main limitations was the way in which the selected method was applied. For example, Zhu et al. [137] stated that their category relevancy-based authority ranking approach needed a more accurate and stable method for parameter selection. Moreover, there is another type of work claiming that more analyses [43] or methods should have been applied. In total, we found eight publications with this limitation.
- **Labelling.** Labelling data comprehensively and efficiently is a widely needed but challenging task [153]. However, manual labelling of an unknown data set for ML is a tedious task for humans [154]. Manual labelling can be necessary if there is no automated data preprocessing system [155] and, by default, it can lead to limitations related to human reliability and possible consequences of human errors or oversights. Another issue is the scalability of manual labelling since it is a hard task when the amount of data is enormous. Moreover, human evaluation may also have biases because different raters may consider different criteria [132]. Finally, the time spent in the process of labelling directly translates into the high costs associated with research projects [156]. In total, we have found two publications facing this limitation.

In conclusion, we can point out that the weakest parts of the selected articles are the lack of data while the methodology could be improved in some studies. However, only 23 publications raised the limitations mentioned above; the rest of the articles mainly discussed future directions, not mentioning weak points.

V. DISCUSSION

In this section, we first present a summary and discussion of our main findings. Next, we provide a discussion that goes beyond those findings. Finally, we will present the implications of our research and the limitations of the selected approach.

A. Key Findings

Here we summarise the main outcomes of our research work, highlighting the most important parts of our survey results.

First of all, four environments, namely, *content sharing & consumption*, *games*, *online learning* and *social networks* emerged from the coding process. Across these environments, we observed measurement and/or development of various capabilities such as *expertise*, *language proficiency* and *soft skills* as well as various behavioural patterns including engagement, influence or dropout, which we grouped as *behaviour*. The most striking result that emerged is that all environments are significantly correlated with all the capabilities stemming from the survey. Further analysis showed that the *content sharing & consumption* environment prevails in the publications. Similarly, strong evidence of the development of expertise was found in this environment.

In an attempt to perform this analysis, we first extracted ways of accessing data in the selected publications, that is, using data published as an *open data set*, using an *API*, using data from the *public domain* or having *direct access* to data. The last two substantially prevailed across the analysed studies. Next, four types of data emerged: *textual*, *clickstream*, *audiovisual* and *biometric*. Remarkably, the vast majority of publications in environments such as *content sharing & consumption* and *social networks* used *textual* data while *video games* and *online learning* environments primarily generated *clickstream* data. After exploring what types of data were found, we aimed to examine the data analysis methods. According to the selected articles, such analysis was conducted through various methods, namely, *ML*, *Network Analysis*, *NLP*, *statistics* and *experimental design*. Our study provided further evidence that all the methods mentioned above have been applied to *textual* data, but that the other types of data are not as flexible in terms of methods. Ultimately, we discussed whether the authors validated the results and what limitations they found in their work. We did not find many studies performing manual validation of their results; however, among those that did, validation of *expertise* finding through the *content sharing & consumption* environment was the most frequent. Lastly, we concluded that the main limitations raised as part of the selected articles could be the lack of data or the methods applied to them.

Fig. 9 shows a summary of the environments and their categorisation with the corresponding data types, their access type and applied methods to validate different competencies and capabilities. We present the results based on the surveyed studies; however, the existence of other types of multimedia environments remains an open question.

B. Extending Beyond the Results

We explored several digital environments that can generate rich data sets through the users' interaction and where data can be used to explicitly or implicitly perform a data-driven evaluation of competencies and capabilities. From our survey, we can extract several general characteristics that environments needed to have for data-driven evaluation of capabilities. First of all, publications within the scope of our survey explored environments that can generate large amounts of data. These data can be accessed either with *direct access* to data or in the *public domain*, but at the same time, it is possible to do it through an *API* or to download an *open data set*. Either way, data of different

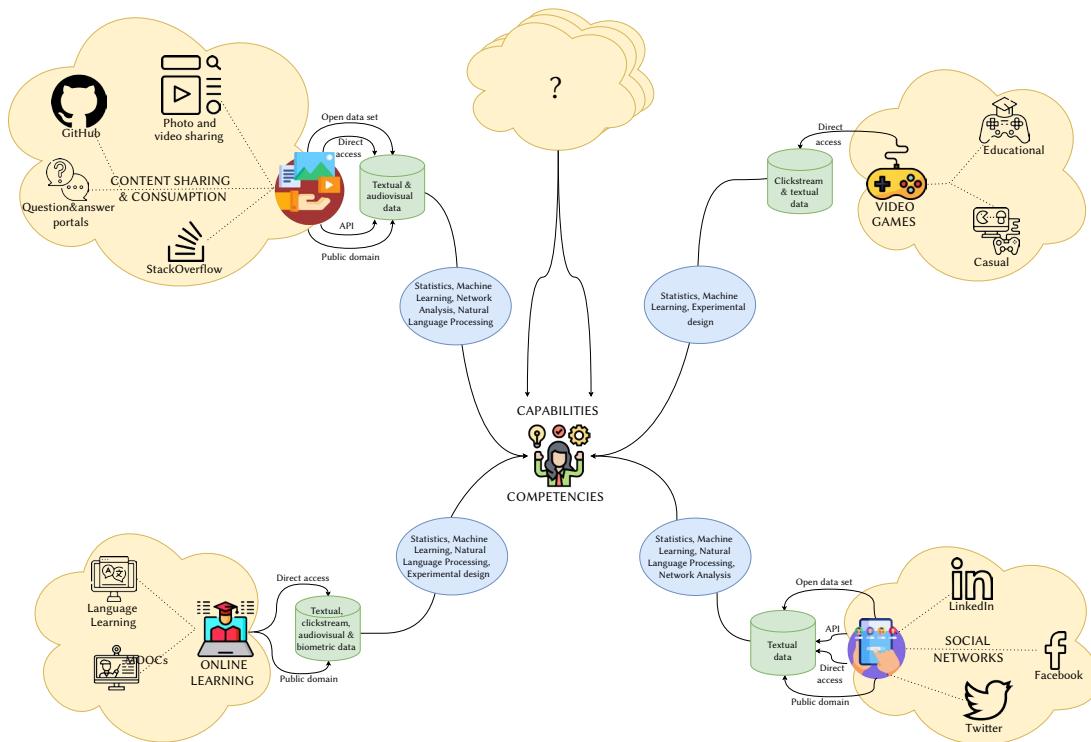


Fig. 9. Overall summary of environments, data access, data type and methods.

types were accessed, where *clickstream* and *textual* data formats prevail over *audiovisual* and *biometric* data. Data access and applied methods are based on the most common types of data; therefore, the rest might prove to be more challenging. Accordingly, other digital environments not found as part of this survey but which fit these criteria also represent potential opportunities to achieve this goal.

As a case in point, we would like to mention several specific environments found in our survey that do not measure capabilities but hold the potential to do so. One of them is Netflix²⁶ – a media streaming platform described in detail in [157]. Unlike publications within the scope of our survey, this study described the Netflix recommendations system, which helps its users make better decisions. Nevertheless, the authors used vast amounts of data that describe what each Netflix user watches, in what way s/he does it (e.g., the device, time of the day, day of the week, the intensity of watching), the place in which each video was discovered and even the recommendations shown but not played in each session. We believe that these data could hold the potential to evaluate various competencies and capabilities such as language proficiency, among others.

Another example is the social payments platform Venmo²⁷, which was one of the most surprising findings of our research. It does not match any of the environments detected during the review process, and it does not fit, a priori, the survey. However, we found one publication [152] aiming to understand users' changes in behaviour over time. Its authors accessed nearly 340 million transactions. Each transaction consists of the transaction ID, sender, receiver, message, time created, amongst other kinds of metadata generated from public posts. Since the authors have already explored users' behaviour, we believe that this portal has the potential for conducting the research according to the goal of our work.

Furthermore, interactive museums turned out to be another unexpected environment that attracted the attention of some

researchers. The authors of [158] measured learners' engagement by using multi-channel data such as eye-tracking, facial expression, posture and interaction logs. These data were captured from visitors' interactions with a fully-instrumented version of a tabletop science exhibit for environmental sustainability called FutureWorlds. We consider it as an environment with ample opportunities for evaluating its users' competencies and capabilities.

C. Implications/Limitations

Our initial objective was to develop a survey on previous work that has performed a data-driven evaluation of competencies across different multimedia environments. The rationale is that the world is shifting towards a focus on capabilities instead of content. Therefore, this issue is of the utmost importance as it will lead society to better adapt to the jobs of the future. Thus, we have foreseen several implications of our results. While the research that we have found shows that it is methodologically feasible to measure competencies based on the data generated in those multimedia environments, it is unclear how to translate these findings into the formal education ecosystem. Some possibilities might include a more frequent utilisation of educational games and other digital environments as part of the classroom activities so that teachers can receive information regarding their students' capabilities to provide personalised feedback. Therefore, more research, products and validations in formal educational settings will be required during the next decade. In that sense, our survey could become a starting point for further research, since it confirms the methodological viability of these approaches. Moreover, it examines new multimedia data-rich environments and their opportunities to support the development of lifelong and lifewide 21st-century capabilities. The current study has some theoretical and practical implications and limitations which will be outlined in this section.

1. Theoretical Implications

Any data set invariably constitutes a biased representation of the population [159]. Moreover, there are unfair practices against

²⁶ <https://netflix.com/>

²⁷ <https://venmo.com/>

members of vulnerable or underrepresented groups, which include the explicit use of protected data attributes such as age or gender, as well as indirect discrimination that occurs when group status is exploited inadvertently [160], [161]. Other biases in data may involve race or ethnicity. The authors of [162], for example, showed that a widely used algorithm, typically used for industry-wide approaches and which is affecting millions of patients, exhibits significant racial bias. Therefore, more work is required for ensuring the fairness and equity of the methods applied in these studies.

2. Practical Implications

We should like to discuss the research challenges and gaps that arise today within this topic. It is essential to mention that we observed a considerable potential based on many studies spanning diverse multimedia environments. We found that every single study applied different methodologies and processes to transform data into capabilities. This means that there is a significant effort invested within this process in the studies. This problem can potentially be improved by proposing a base framework that can be applied to infer capabilities based on data while generalising well across different digital environments; to the best of our knowledge, this kind of framework does not yet exist. Accordingly, we believe that our work can serve as a knowledge basis for building it. However, we have detected that there are not many examples of using *audiovisual* or *biometric* data in the studies that we explored. This could lead to having insufficient evidence of the development of methods that use those data types to measure some capabilities across various environments. Moreover, easy identification of an individual with very few data points leads to the restricted ethical use of data and their purposes. Therefore, while having the potential to be useful, it might not be possible to use some data given the aforementioned issue and the fact that users' privacy must be respected.

3. Limitations

The most important limitation of this work that could influence the obtained results lies in the search process. This is because the present study only investigated the environments in which we could find evidence of measuring competencies and capabilities that emerged from the coding process (see Section D). Even though we covered the most relevant environments in the context of the data-driven evaluation of competencies and capabilities, there could be other environments we are not aware of, and consequently, they are not included in our study. Despite this fact, we assume that we found evidence of capabilities evaluation across all the environments discussed throughout this survey.

4. Novel Contributions

The strength of the current paper is that we identified and reviewed studies that have been able to use different types of data and analyses to infer a range of competencies and capabilities in the four multimedia environments that emerged as part of the survey. As we mentioned in Section II, all previous studies explored how to develop and measure competencies based on only one specific type of online platform. Our work has provided a more complete picture of the multiple existing multimedia environments that can be used for evaluating different competencies and capabilities. What we learned in the survey is a key starting point for the potential change in the educational and training systems, suggesting new data-driven assessment possibilities that represent new steps forward to provide personalised feedback. At the same time, our work may motivate other researchers to perform additional experiments to learn of new digital environments holding the potential to measure and/or develop various capabilities.

VI. CONCLUSIONS AND FUTURE WORK

This work represents a groundbreaking analysis of current literature examining diverse technology-mediated environments that can generate rich data sets through the users' interaction and where data can be used to perform a data-driven evaluation of competencies and capabilities. This is the first time, as far as we know, that this kind of research was conducted. Despite facing an ill-defined area, this study deeply enhanced our current understanding of this open research line. In this regard, we provided an overview of the existing research as well as concluded that all the environments we discussed (*content sharing & consumption, video games, online learning and social networks*) proved their ability to generate rich data sets through the users' interaction. We found evidence that all these environments are highly correlated with the measurement and/or development of various capabilities such as *expertise, language proficiency and soft skills*. According to the over one hundred surveyed studies, this measurement was done with the application of different methods (*ML, Network Analysis, NLP, statistics and experimental design*), which we also discussed in detail.

We believe that our survey encompasses numerous new approaches that confirm the viability of performing data-driven evaluations of competencies and capabilities. We are confident that based on our results, it is possible to develop a framework that can generalise well to different environments, data types and capabilities, and that this can help to conduct additional research by re-applying the framework in future studies. Accordingly, more research is needed to be able to transfer these ideas into formal education settings with new innovative products. In the future, teachers will be able to use such products to better assess the capabilities of their students and to provide personalised feedback. On the other hand, there is a need to develop this research evaluating the algorithmic bias issues as well as being respectful of students' privacy. Our future work will focus on exploring several of these multimedia environments with the aim of developing our own algorithms for measuring the capabilities. More specifically, we will focus on the measurement of expertise and soft skills across different environments, trying to analyse different types of data by applying various methods.

APPENDIX

A. Acronyms

Acronym	Reference abbreviation
MOOC	Massive Open Online Course
Q&A	Question-and-Answer format
CQA	Community Question Answering
RQ	Research Questions
OVP	Online Video Platform
API	Application Programming Interface
HRV	Heart Rate Variability
ANOVA	Analysis of Variance
ML	Machine Learning
NN	Neural Network
SVM	Support Vector Machine
HITS	Hyperlink-Induced Topic Search
NLP	Natural Language Processing
EDA	Electrodermal activity
PCA	Principal Component Analysis
N/A	Not Applicable
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses

B. Results of the Coding Process

Title	Environment (RQ1)	Data access (RQ2)	Data (RQ3)	Methods (RQ4)	Skills (RQ5)	Validation (RQ6)
[134]	CSC ¹	Public domain	Text	Network analysis	Expertise	No
[49]	CSC	Public domain	Text	ML	Expertise	No
[54]	CSC	Public domain	Text	Network analysis	Expertise	No
[57]	CSC	N/A	Text	Statistics	Expertise	No
[151]	CSC	Public domain	Text	ML	Expertise	Yes
[129]	CSC	Open data set	Text	ML	Expertise	No
[103]	CSC	Open data set	Text	Statistics	Expertise	No
[133]	CSC	Public domain	Text	Statistics	Expertise	No
[122]	CSC	Public domain	Text	Network analysis	Expertise	No
[128]	CSC	Open data set	Text	Statistics	Expertise	No
[138]	CSC	Public domain	Text	ML	Expertise	Yes
[55]	CSC	N/A	Text	Statistics	Expertise	No
[135]	CSC	Public domain	Text	ML	Expertise	No
[39]	CSC	N/A	Text	Network analysis, NLP	Expertise	No
[60]	CSC	API	Text	N/A	Expertise	Yes
[61]	CSC	API	Text	Statistics	Expertise	Yes
[121]	CSC	Public domain	Text	Network analysis	Expertise	Yes
[126]	CSC	API	Text	ML	Expertise	Yes
[136]	CSC	Public domain	Text	ML	Expertise	No
[132]	CSC	Public domain	Text	Network analysis	Expertise	Yes
[59]	CSC	Public domain	Text	ML	Expertise	No
[125]	CSC	API	Text	Statistics	Expertise	Yes
[137]	CSC	Public domain	Text	Network analysis	Expertise	Yes
[53]	CSC	API	Text	Network analysis	Expertise	Yes
[109]	CSC	Public domain	Text	Network analysis	Expertise	Yes
[50]	CSC	API	Text	NLP	Behaviour	No
[97]	CSC	API	Text	ML	N/A	No
[58]	CSC	Public domain	Text	Statistics	N/A	No
[46]	CSC	Open data set	Text	Statistics	Behaviour	No
[56]	CSC	Public domain	Audiovisual	Statistics	Expertise	No
[108]	CSC	Public domain	Text	Statistics	Expertise	No
[42]	CSC	N/A	Text	Statistics	Behaviour	No
[127]	CSC	Public domain	Text	NLP	N/A	No
[99]	CSC	Public domain	Text	Statistics	Behaviour	No
[51]	CSC	Open data set	Text	Statistics	Behaviour	No
[102]	CSC	Open data set	Text	Statistics	Behaviour	No
[157]	CSC	Direct access	Text	ML	N/A	No
[101]	CSC	API	Text	N/A	N/A	N/A
[141]	CSC	Direct access	Text	Statistics	Language ²	No
[40]	CSC	Open data set	Text	Network analysis, ML	Behaviour	No
[130]	CSC	Open data set	Text	Statistics	Expertise	No
[93]	Video games	Direct access	Clickstream	ML	Expertise	No
[69]	Video games	Direct access	Clickstream	ML	Expertise	No
[71]	Video games	Direct access	Clickstream	Statistics	Behaviour	No
[65]	Video games	Direct access	Clickstream	ML	Expertise	Yes
[70]	Video games	Direct access	Clickstream	Statistics	Behaviour	No
[66]	Video games	Direct access	Clickstream	ML	Expertise	No
[146]	Video games	Direct access	Clickstream	Statistics, Experiment ³	Soft skills	No
[147]	Video games	Direct access	Clickstream	Experiment	Soft skills	No
[149]	Video games	Direct access	Clickstream	Experiment	Soft skills	Yes
[155]	Video games	Direct access	Text	Statistics	Behaviour	No
[64]	Video games	Direct access	Text	Experiment	Expertise	No

Title	Environment (RQ1)	Data access (RQ2)	Data (RQ3)	Methods (RQ4)	Skills (RQ5)	Validation (RQ6)
[63]	Video games	Direct access	Clickstream	ML	Behaviour	No
[112]	Video games	Direct access	Clickstream	Statistics	Behaviour	No
[100]	Online Learning	Public domain	Text	ML	Soft skills	No
[98]	Online Learning	Public domain	Clickstream	Statistics	Expertise	No
[78]	Online Learning	Public domain	Text	Statistics	Language	No
[82]	Online Learning	Direct access	Biometric	ML	Expertise	No
[81]	Online Learning	Direct access	Audiovisual	ML	Soft skills	No
[117]	Online Learning	Direct access	Biometric	ML	Soft skills	No
[73]	Online Learning	N/A	Clickstream	NLP	Expertise	No
[143]	Online Learning	Direct access	Audiovisual	ML	Language	No
[142]	Online Learning	Direct access	Text	Experiment	Language	Yes
[110]	Online Learning	Direct access	Text	Statistics, Experiment	Language	No
[83]	Online Learning	Direct access	Text	Statistics	Behaviour	No
[76]	Online Learning	Direct access	Clickstream	ML	Behaviour	Yes
[139]	Online Learning	Direct access	Clickstream	Statistics	Expertise	No
[74]	Online Learning	Direct access	Audiovisual	NLP	Soft skills	No
[114]	Online Learning	Direct access	Clickstream	Statistics	Soft skills	No
[75]	Online Learning	N/A	Text	Statistics	Expertise	No
[116]	Online Learning	Direct access	Audiovisual	ML	Expertise	No
[48]	Online Learning	Direct access	Text	NLP	Behaviour	No
[106]	Online Learning	Public domain	Text	Network analysis	Behaviour	No
[158]	Online Learning	Direct access	Biometric	ML	Behaviour	No
[111]	Online Learning	Direct access	Text	Statistics	Language	No
[47]	Online Learning	Public domain	Clickstream	ML	Behaviour	No
[120]	Online Learning	Public domain	Clickstream	ML	Expertise	No
[118]	Online Learning	Direct access	Clickstream	Statistics	Behaviour	No
[84]	Online Learning	Direct access	Clickstream	Experiment	Behaviour	No
[77]	Online Learning	Public domain	Clickstream	Statistics	Behaviour	No
[113]	Online Learning	Direct access	Clickstream	Statistics	Behaviour	No
[119]	Online Learning	Direct access	Text	Statistics	Expertise	No
[80]	Online Learning	Direct access	Text	Statistics	N/A	No
[79]	Online Learning	Direct access	Text	Statistics, Experiment	N/A	No
[94]	Online Learning	Direct access	Text	Statistics, Experiment	Behaviour	No
[52]	Social Network	Public domain	Text	Statistics	Expertise	No
[95]	Social Network	API	Text	ML	Expertise	Yes
[107]	Social Network	Direct access	Text	ML	Expertise	Yes
[140]	Social Network	API	Text	ML	Language	No
[41]	Social Network	Public domain, API	Text	ML	Expertise	No
[96]	Social Network	API	Text	ML	Soft skills	No
[89]	Social Network	Direct access	Text	ML	Expertise	Yes
[88]	Social Network	Public domain	Text	Statistics	Language	No
[105]	Social Network	API	Text	ML	Behaviour	No
[45]	Social Network	API	Text	Statistics	Behaviour	No
[87]	Social Network	Direct access	Text	Statistics, Experiment	Language	Yes
[43]	Social Network	API	Text	NLP	Behaviour	No
[44]	Social Network	API	Text	Network analysis	Behaviour	No
[91]	Social Network	Direct access	Text	Statistics	Expertise	No
[104]	Social Network	Public domain	Text	NLP	N/A	No
[92]	Social Network	Direct access	Text	Statistics	Soft skills	No
[152]	N/A	API	Text	Statistics	Behaviour	No

¹ Content sharing & consumption

² Language proficiency

³ Experimental design

ACKNOWLEDGMENT

This study was partially funded by the Spanish Government grants IJC2020-044852-I and RYC-2015-18210, co-funded by the European Social Fund, as well as by the COBRA project (10032/20/0035/00), granted by the Spanish Ministry of Defence and by the SCORPION project (21661-PDC-21), granted by the Seneca Foundation of the Region of Murcia, Spain.

REFERENCES

- [1] B. S. Bloom, "The new direction in educational research: Alterable variables," *The Journal of Negro Education*, vol. 49, no. 3, pp. 337–349, 1980.
- [2] C. R. Wolfe, "Creating informal learning environments on the world wide web," in *Learning and teaching on the World Wide Web*, Elsevier, 2001, pp. 91–112.
- [3] S. Downes, et al., "New technology supporting informal learning," *Journal of emerging technologies in web intelligence*, vol. 2, no. 1, pp. 27–33, 2010.
- [4] A. Nandi, M. Mandernach, "Hackathons as an informal learning platform," in *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, SIGCSE '16, New York, NY, USA, 2016, p. 346–351, Association for Computing Machinery.
- [5] J. Maldonado-Mahauad, M. Pérez-Sanagustín, R. F. Kizilcec, N. Morales, J. Munoz-Gama, "Mining theory-based patterns from big data: Identifying self-regulated learning strategies in massive open online courses," *Computers in Human Behavior*, vol. 80, pp. 179 – 196, 2018, doi: <https://doi.org/10.1016/j.chb.2017.11.011>.
- [6] M. Anshari, Y. Alas, L. S. Guan, "Developing online learning resources: Big data, social networks, and cloud computing to support pervasive knowledge," *Education and Information Technologies*, vol. 21, no. 6, pp. 1663–1677, 2016.
- [7] R. Kizilcec, C. Brooks, "Diverse Big Data and Randomized Field Experiments in Massive Open Online Courses," in *The Handbook of Learning Analytics*, C. Lang, G. Siemens, A. F. Wise, D. Gašević Eds., Alberta, Canada: Society for Learning Analytics Research (SoLAR), 2017, pp. 211–222, 1 ed.
- [8] R. Eynon, "The rise of big data: what does it mean for education, technology, and media research?," *Learning, Media and Technology*, vol. 38, no. 3, pp. 237– 240, 2013, doi: [10.1080/17439884.2013.771783](https://doi.org/10.1080/17439884.2013.771783).
- [9] J.-E. Mai, "Big data privacy: The datafication of personal information," *The Information Society*, vol. 32, no. 3, pp. 192–199, 2016.
- [10] V. Mayer-Schönberger, K. Cukier, *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.
- [11] J. Van Dijck, "Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology," *Surveillance & Society*, vol. 12, no. 2, pp. 197–208, 2014.
- [12] H.-U. Otto, H. Ziegler, "Capabilities and education," *Social Work & Society*, vol. 4, no. 2, pp. 269–287, 2006.
- [13] D. Kember, D. Y. Leung, R. S. Ma, "Characterizing learning environments capable of nurturing generic capabilities in higher education," *Research in Higher Education*, vol. 48, no. 5, p. 609, 2007.
- [14] M. Pinzone, P. Fantini, S. Perini, S. Garavaglia, M. Taisch, G. Miragliotta, "Jobs and skills in industry 4.0: An exploratory research," in *Advances in Production Management Systems. The Path to Intelligent, Collaborative and Sustainable Manufacturing*, Cham, 2017, pp. 282–288, Springer International Publishing.
- [15] A. Smith, J. Anderson, "AI, robotics, and the future of jobs," *Pew Research Center*, vol. 6, p. 51, 2014.
- [16] N. P. Stromquist, *Education in a globalized world: The connectivity of economic power, technology, and knowledge*. Rowman & Littlefield, 2002.
- [17] C. Redecker, M. Leis, M. Leendertse, Y. Punie, G. Gijssbers, P. Kirschner, S. Stoyanov, B. Hoogveld, "The future of learning: New ways to learn new skills for future jobs," *Results from an online expert consultation. Technical Note JRC60869, JRC-IPTS, Seville*, 2010.
- [18] J. C.-Y. Sun, R. Rueda, "Situational interest, computer self-efficacy and self-regulation: Their impact on student engagement in distance education," *British Journal of educational technology*, vol. 43, no. 2, pp. 191– 204, 2012.
- [19] R. Nagarajan, R. Prabhu, "Competence and capability: A new look," *International Journal of Management*, vol. 6, no. 6, pp. 7–11, 2015.
- [20] R. Hipkins, "Competencies or capabilities," *He Whakaaro An, Se2*, vol. 3, pp. 55–57, 2013.
- [21] R. E. Boyatzis, *The competent manager: A model for effective performance*. John Wiley & Sons, 1982.
- [22] R. Henderson, I. Cockburn, "Measuring competence? exploring firm effects in pharmaceutical research," *Strategic management journal*, vol. 15, no. S1, pp. 63–84, 1994.
- [23] J. F. Lozano, A. Boni, J. Peris, A. Hueso, "Competencies in higher education: A critical analysis from the capabilities approach," *Journal of Philosophy of Education*, vol. 46, no. 1, pp. 132–147, 2012.
- [24] V. Vathanophas, "Competency requirements for effective job performance in Thai public sector," *Contemporary management research*, vol. 3, no. 1, p. 45, 2007.
- [25] J. Macnamara, "Competence, competencies and/or capabilities for public communication? a public sector study," *Asia Pacific Public Relations Journal*, vol. 19, pp. 16–40, 2018.
- [26] P. Morgan, "The concept of capacity," *European Centre for Development Policy Management*, pp. 1–19, 2006.
- [27] L. Vincent, "Differentiating competence, capability and capacity," *Innovating Perspectives*, vol. 16, no. 3, pp. 1–2, 2008.
- [28] M. Z. Al-Taie, S. Kadry, A. I. Obasa, "Understanding expert finding systems: domains and techniques," *Social Network Analysis and Mining*, vol. 8, no. 1, p. 57, 2018.
- [29] O. Husain, N. Salim, R. A. Alias, S. Abdelsalam, A. Hassan, "Expert finding systems: A systematic review," *Applied Sciences*, vol. 9, no. 20, p. 4250, 2019.
- [30] I. Srba, M. Bielikova, "A comprehensive survey and classification of approaches for community question answering," *ACM Transactions on the Web (TWEB)*, vol. 10, no. 3, pp. 1–63, 2016.
- [31] 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, no. 2, pp. 661 – 686, 2012, doi: <https://doi.org/10.1016/j.compedu.2012.03.004>.
- [32] E. A. Boyle, T. Hainey, T. M. Connolly, G. Gray, J. Earp, M. Ott, T. Lim, M. Ninaus, C. Ribeiro, J. Pereira, "An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games," *Computers & Education*, vol. 94, pp. 178 – 192, 2016, doi: <https://doi.org/10.1016/j.compedu.2015.11.003>.
- [33] T. Hainey, T. M. Connolly, E. A. Boyle, A. Wilson, A. Razak, "A systematic literature review of games- based learning empirical evidence in primary education," *Computers & Education*, vol. 102, pp. 202– 223, 2016.
- [34] X. Wei, N. Saab, W. Admiraal, "Assessment of cognitive, behavioral, and affective learning outcomes in massive open online courses: A systematic literature review," *Computers & Education*, vol. 163, p. 104097, 2021, doi: <https://doi.org/10.1016/j.compedu.2020.104097>.
- [35] M. J. Page, D. Moher, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, et al., "PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews," *BMJ*, vol. 372, 2021.
- [36] R. Huelin, I. Iheanacho, K. Payne, K. Sandman, "What's in a name? systematic and non-systematic literature reviews, and why the distinction matters," *The evidence*, pp. 34–37, 2015.
- [37] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, EASE '14, New York, NY, USA, 2014, Association for Computing Machinery.
- [38] J. Randolph, "A guide to writing the dissertation literature review," *Practical Assessment, Research, and Evaluation*, vol. 14, no. 1, p. 13, 2009.
- [39] F. Riahi, Z. Zolaktaf, M. Shafiei, E. Milios, "Finding expert users in community question answering," in *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, New York, NY, USA, 2012, p. 791–798, Association for Computing Machinery.
- [40] D. Movshovitz-Attias, Y. Movshovitz-Attias, P. Steenkiste, C. Faloutsos, "Analysis of the reputation system and user contributions on a question answering website: StackOverflow," in *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, 2013, pp. 886–893, IEEE.

- [41] Y. Xu, D. Zhou, S. Lawless, "Inferring your expertise from Twitter: Integrating sentiment and topic relatedness," in *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 2016, pp. 121–128, IEEE.
- [42] L. Akritidis, D. Katsaros, P. Bozaris, "Identifying the productive and influential bloggers in a community," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 5, pp. 759–764, 2011.
- [43] Y. Bae, H. Lee, "Sentiment analysis of Twitter audiences: Measuring the positive or negative influence of popular twitterers," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 12, pp. 2521–2535, 2012.
- [44] C. Bigonha, T. N. Cardoso, M. M. Moro, M. A. Gonçalves, V. A. Almeida, "Sentiment-based influence detection on Twitter," *Journal of the Brazilian Computer Society*, vol. 18, no. 3, pp. 169–183, 2012.
- [45] M. Cha, H. Haddadi, F. Benevenuto, P. K. Gummadi, et al., "Measuring user influence in Twitter: The million follower fallacy," *Icwsn*, vol. 10, no. 10-17, p. 30, 2010.
- [46] H. U. Khan, A. Daud, "Finding the top influential bloggers based on productivity and popularity features," *New Review of Hypermedia and Multimedia*, vol. 23, no. 3, pp. 189–206, 2017.
- [47] S. Nagrecha, J. Z. Dillon, N. V. Chawla, "MOOC dropout prediction: Lessons learned from making pipelines interpretable," in *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, Republic and Canton of Geneva, CHE, 2017, p. 351–359, International World Wide Web Conferences Steering Committee.
- [48] N. Dmoshinskaia, "Dropout prediction in MOOCs: using sentiment analysis of users' comments to predict engagement," Master's thesis, University of Twente, 2016.
- [49] S. Patil, K. Lee, "Detecting experts on Quora: by their activity, quality of answers, linguistic characteristics and temporal behaviors," *Social network analysis and mining*, vol. 6, no. 1, p. 5, 2016.
- [50] D. Choi, J. Han, T. Chung, Y.-Y. Ahn, B.-G. Chun, T. T. Kwon, "Characterizing conversation patterns in Reddit: From the perspectives of content properties and user participation behaviors," in *Proceedings of the 2015 ACM on Conference on Online Social Networks*, COSN '15, New York, NY, USA, 2015, p. 233–243, Association for Computing Machinery.
- [51] J. Yang, K. Tao, A. Bozzon, G.-J. Houben, "Sparrows and owls: Characterisation of expert behaviour in StackOverflow," in *User Modeling, Adaptation, and Personalization*, Cham, 2014, pp. 266–277, Springer, Springer International Publishing.
- [52] E. Malherbe, M.-A. Aufaure, "Bridge the terminology gap between recruiters and candidates: A multilingual skills base built from social media and linked data," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016, pp. 583–590, IEEE.
- [53] G. Zhou, S. Lai, K. Liu, J. Zhao, "Topic-sensitive probabilistic model for expert finding in question answer communities," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, New York, NY, USA, 2012, p. 1662–1666, Association for Computing Machinery.
- [54] P. Jurczyk, E. Agichtein, "Discovering authorities in question answer communities by using link analysis," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007, pp. 919–922.
- [55] N. Raj, L. Dey, B. Gaonkar, "Expertise prediction for social network platforms to encourage knowledge sharing," in *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '11, USA, 2011, p. 380–383, IEEE Computer Society.
- [56] A. Raikos, P. Waidyasekara, "How useful is YouTube in learning heart anatomy?," *Anatomical sciences education*, vol. 7, no. 1, pp. 12–18, 2014.
- [57] A. Pal, A. Herdagdelen, S. Chatterji, S. Taank, Chakrabarti, "Discovery of topical authorities in Instagram," in *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, Republic and Canton of Geneva, CHE, 2016, p. 1203–1213, International World Wide Web Conferences Steering Committee.
- [58] J. T. Hertel, N. M. Wessman-Enzinger, "Examining Pinterest as a curriculum resource for negative integers: An initial investigation," *Education Sciences*, vol. 7, no. 2, p. 45, 2017.
- [59] A.-M. Popescu, K. Y. Kamath, J. Caverlee, "Mining potential domain expertise in Pinterest," in *UMAP workshops*, 2013.
- [60] J. Oliveira, M. Vigiato, E. Figueiredo, "How well do you know this library? mining experts from source code analysis," in *Proceedings of the XVIII Brazilian Symposium on Software Quality*, SBQS'19, New York, NY, USA, 2019, p. 49–58, Association for Computing Machinery.
- [61] R. Saxena, N. Pedanekar, "I know what you coded last summer: Mining candidate expertise from GitHub repositories," in *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17 Companion, New York, NY, USA, 2017, p. 299–302, Association for Computing Machinery.
- [62] P. A. Martínez, M. J. Gómez, J. A. Ruipérez-Valiente, G. Martínez Pérez, Y. J. Kim, "Visualizing educational game data: A case study of visualizations to support teachers," in *Learning Analytics Summer Institute Spain 2020: Learning Analytics. Time for Adoption?*, Jun 2020, pp. 97–111, CEUR Workshop Proceedings.
- [63] E. Harpstead, T. Zimmermann, N. Nagapan, J. J. Guajardo, R. Cooper, T. Solberg, D. Greenawalt, "What drives people: Creating engagement profiles of players from game log data," in *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '15, New York, NY, USA, 2015, p. 369–379, Association for Computing Machinery.
- [64] A. J. Lesser, *Video game technology and learning in the music classroom*. PhD dissertation, Teachers College, Columbia University, 2019.
- [65] P. Kantharaju, K. Alderfer, J. Zhu, B. Char, B. Smith, S. Ontañón, "Tracing player knowledge in a parallel programming educational game," 2019.
- [66] F. Chen, Y. Cui, M.-W. Chu, "Utilizing game analytics to inform and validate digital game-based assessment with evidence-centered game design: A case study," *International Journal of Artificial Intelligence in Education*, vol. 30, no. 3, pp. 481–503, 2020.
- [67] J. A. Ruipérez-Valiente, M. Gaydos, L. Rosenheck, Y. J. Kim, E. Klopfer, "Patterns of engagement in an educational massive multiplayer online game: A multidimensional view," *IEEE Transactions on Learning Technologies*, 2020.
- [68] Y. J. Kim, J. A. Ruipérez-Valiente, "Data-driven game design: The case of difficulty in educational games," in *European Conference on Technology Enhanced Learning*, 2020, pp. 449–454, Springer.
- [69] M.-T. Cheng, Y.-W. Lin, H.-C. She, "Learning through playing Virtual Age: Exploring the interactions among student concept learning, gaming performance, in-game behaviors, and the use of in-game characters," *Computers & Education*, vol. 86, pp. 18–29, 2015.
- [70] J. Kang, M. Liu, W. Qu, "Using gameplay data to examine learning behavior patterns in a serious game," *Computers in Human Behavior*, vol. 72, pp. 757–770, 2017.
- [71] W. Westera, R. Nadolski, H. Hummel, "Serious gaming analytics: What students' log files tell us about gaming and learning," *International Journal of Serious Games*, vol. 1, Jun. 2014, doi: 10.17083/ijsg.v1i2.9.
- [72] T. Daradoumis, R. Bassi, F. Xhafa, S. Caballé, "A review on massive e-learning (MOOC) design, delivery and assessment," in *2013 Eighth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, 2013, pp. 208–213.
- [73] S. Crossley, L. Paquette, M. Dascalu, D. S. McNamara, R. S. Baker, "Combining click-stream data with NLP tools to better understand MOOC completion," in *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, LAK '16, New York, NY, USA, 2016, p. 6–14, Association for Computing Machinery.
- [74] R. Zhao, V. Li, H. Barbosa, G. Ghoshal, M. E. Hoque, "Semi-automated 8 collaborative online training module for improving communication skills," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 2, pp. 1–20, 2017.
- [75] R. Reddick, "Using a Glicko-based algorithm to measure in-course learning," *International Educational Data Mining Society*, 2019.
- [76] X. Wang, D. Yang, M. Wen, K. Koedinger, C. P. Rosé, "Investigating how student's cognitive behavior in MOOC discussion forums affect learning gains," *International Educational Data Mining Society*, 2015.
- [77] J. Kim, P. J. Guo, D. T. Seaton, P. Mitros, K. Z. Gajos, R. C. Miller, "Understanding in-video dropouts and interaction peaks in online lecture videos," in *Proceedings of the First ACM Conference on Learning @ Scale Conference*, L@S '14, New York, NY, USA, 2014, p. 31–40, Association for Computing Machinery.
- [78] D. Huynh, L. Zuo, H. Iida, "Analyzing gamification of "Duolingo" with focus on its course structure," in *International Conference on Games and Learning Alliance*, 2016, pp. 268–277, Springer.

- [79] S. Chootongchai, N. Songkram, "Design and development of seci and moodle online learning systems to enhance thinking and innovation skills for higher education learners," *International Journal of Emerging Technologies in Learning (IJET)*, vol. 13, no. 03, pp. 154–172, 2018.
- [80] B. Hightower, C. Rawl, M. Schutt, "Collaborations for delivering the library to students through WebCT," *Reference Services Review*, 2007.
- [81] H. Tanaka, S. Sakti, G. Neubig, T. Toda, H. Negoro, H. Iwasaka, S. Nakamura, "Automated social skills trainer," in *Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI '15*, New York, NY, USA, 2015, p. 17–27, Association for Computing Machinery.
- [82] P. Pham, J. Wang, "Attentivelearner: Improving mobile MOOC learning via implicit heart rate tracking," in *Artificial Intelligence in Education*, Cham, 2015, pp. 367–376, Springer, Springer International Publishing.
- [83] C. Yin, F. Okubo, A. Shimada, M. Oi, S. Hirokawa, H. Ogata, "Identifying and analyzing the learning behaviors of students using e-books," in *Doctoral Student Consortium (DSC) - Proceedings of the 23rd International Conference on Computers in Education, ICCE 2015*, 2015, pp. 118–120, Asia-Pacific Society for Computers in Education.
- [84] A. Shimada, K. Mouri, H. Ogata, "Real-time learning analytics of e-book operation logs for on-site lecture support," in *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*, 2017, pp. 274–275, IEEE.
- [85] D. M. Boyd, N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of computer-mediated communication*, vol. 13, no. 1, pp. 210–230, 2007.
- [86] J. Pastor-Galindo, M. Zago, P. Nespoli, S. López Bernal, A. Huertas, M. Pérez, J. A. Ruipérez-Valiente, G. Martínez Pérez, F. Gómez Mármol, "Spotting political social bots in Twitter: A use case of the 2019 spanish general election," *IEEE Transactions on Network and Service Management*, vol. 17, pp. 2156–2170, 10 2020, doi: 10.1109/TNSM.2020.3031573.
- [87] E. Özdemir, "Promoting EFL learners' intercultural communication effectiveness: a focus on Facebook," *Computer Assisted Language Learning*, vol. 30, no. 6, pp. 510–528, 2017.
- [88] W. Orawiwatnakul, S. Wichadee, "Achieving better learning performance through the discussion activity in Facebook," *Turkish Online Journal of Educational Technology-TOJET*, vol. 15, no. 3, pp. 1–8, 2016.
- [89] X. Yan, J. Yang, M. Obukhov, L. Zhu, J. Bai, S. Wu, Q. He, "Social skill validation at LinkedIn," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19*, New York, NY, USA, 2019, p. 2943–2951, Association for Computing Machinery.
- [90] B. Muros-Ruiz, Y. Aragón-Carretero, A. Bustos-Jiménez, "Youth's usage of leisure time with video games and social networks," *Comunicar: Revista Científica de Comunicación y Educación*, vol. 20, no. 40, pp. 31–39, 2013.
- [91] M. Boukessa, L. B. Romdhane, "Identifying authorities in online communities," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 3, pp. 1–23, 2015.
- [92] T. P. Alloway, R. G. Alloway, "The impact of engagement with social networking sites (SNSs) on cognitive skills," *Computers in Human Behavior*, vol. 28, no. 5, pp. 1748–1754, 2012.
- [93] M. Niemelä, T. Kärkkäinen, S. Äyrämö, M. Ronimus, U. Richardson, H. Lyytinen, "Game learning analytics for understanding reading skills in transparent writing system," *British Journal of Educational Technology*, 02 2020, doi: 10.1111/bjet.12916.
- [94] A. I. Wang, A. Lieberoth, "The effect of points and audio on concentration, engagement, enjoyment, learning, motivation, and classroom dynamics using Kahoot!," in *European Conference on Games Based Learning*, vol. 20, 2016, Academic Conferences International Limited.
- [95] M. Bouguessa, L. B. Romdhane, "Identifying authorities in online communities," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 3, pp. 1–23, 2015.
- [96] G. Guerrero, E. Sarchi, F. Tapia, "Predict the personality of Facebook profiles using automatic learning techniques and BFI test," in *New Knowledge in Information Systems and Technologies*, Cham, 2019, pp. 482–493, Springer International Publishing.
- [97] R. Nielek, O. Jarczyk, K. Pawlak, L. Bukowski, R. Bartusiak, A. Wierzbicki, "Choose a job you love: predicting choices of GitHub developers," in *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 2016, pp. 200–207, IEEE.
- [98] A. Cohen, U. Shimony, R. Nachmias, T. Soffer, "Active learners' characterization in MOOC forums and their generated knowledge," *British Journal of Educational Technology*, vol. 50, no. 1, pp. 177–198, 2019.
- [99] J. Han, D. Choi, A.-Y. Choi, J. Choi, T. Chung, T. T. Kwon, J.-Y. Rha, C.-N. Chuah, "Sharing topics in Pinterest: Understanding content creation and diffusion behaviors," in *Proceedings of the 2015 ACM on Conference on Online Social Networks, COSN '15*, New York, NY, USA, 2015, p. 245–255, Association for Computing Machinery.
- [100] F. Calefato, F. Lanubile, B. Vasilescu, "A large-scale, in-depth analysis of developers' personalities in the apache ecosystem," *Information and Software Technology*, vol. 114, pp. 1–20, 2019.
- [101] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, J. Blackburn, "The Pushshift Reddit dataset," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 830–839.
- [102] B. Vasilescu, V. Filkov, A. Serebrenik, "StackOverflow and GitHub: Associations between software development and crowdsourced knowledge," in *2013 International Conference on Social Computing*, 2013, pp. 188–195, IEEE.
- [103] W. H. Lim, M. J. Carman, S.-M. J. Wong, "Estimating relative user expertise for content quality prediction on Reddit," in *Proceedings of the 28th ACM Conference on Hypertext and Social Media, HT '17*, New York, NY, USA, 2017, p. 55–64, Association for Computing Machinery.
- [104] C. Mhamdi, M. Al-Emran, S. A. Salloum, *Text Mining and Analytics: A Case Study from News Channels Posts on Facebook*, pp. 399–415. Cham: Springer International Publishing, 2018.
- [105] W. Xing, F. Gao, "Exploring the relationship between online discourse and commitment in Twitter professional learning communities," *Computers & Education*, vol. 126, pp. 388–398, 2018.
- [106] T. Hecking, I.-A. Chounta, H. U. Hoppe, "Investigating social and semantic user roles in MOOC discussion forums," in *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, LAK '16*, New York, NY, USA, 2016, p. 198–207, Association for Computing Machinery.
- [107] A. Pal, S. Counts, "Identifying topical authorities in microblogs," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM '11*, New York, NY, USA, 2011, p. 45–54, Association for Computing Machinery.
- [108] Z.-J. Yang, J. Lin, Y.-S. Yang, "Identification of network behavioral characteristics of high-expertise users in interactive innovation: The case of forum automobile," *Asia Pacific Management Review*, 2020, doi: https://doi.org/10.1016/j.apmr.2020.06.002.
- [109] H. Zhu, H. Cao, H. Xiong, E. Chen, J. Tian, "Towards expert finding by leveraging relevant categories in authority ranking," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 2221–2224.
- [110] C. MACLEOD, "Evaluating student use of Duolingo, an online self-study platform," *JOURNAL OF ATOMI UNIVERSITY FACULTY OF LITERATURE*, no. 54, pp. A49–A67, 2019.
- [111] S. Loewen, D. Crowther, D. R. Isbell, K. M. Kim, J. Maloney, Z. F. Miller, H. Rawal, "Mobile-assisted language learning: A Duolingo case study," *ReCALL*, vol. 31, no. 3, pp. 293–311, 2019.
- [112] M. Liu, J. Lee, J. Kang, S. Liu, "What we can learn from the data: A multiple-case study examining behavior patterns by students with different characteristics in using a serious game," *Technology, Knowledge and Learning*, vol. 21, no. 1, pp. 33–57, 2016.
- [113] Y. Sun, C. Xin, "Using Coursera clickstream data to improve online education for software engineering," in *Proceedings of the ACM Turing 50th Celebration Conference - China, ACM TUR-C '17*, New York, NY, USA, 2017, Association for Computing Machinery.
- [114] L. Breslow, D. E. Pritchard, J. DeBoer, G. S. Stump, A. D. Ho, D. T. Seaton, "Studying learning in the worldwide classroom research into edX's first MOOC," *Research & Practice in Assessment*, vol. 8, pp. 13–25, 2013.
- [115] G. Alexandron, J. A. Ruipérez-Valiente, Z. Chen, P. J. Muñoz-Merino, D. E. Pritchard, "Copying@Scale: Using harvesting accounts for collecting correct answers in a MOOC," *Computers & Education*, vol. 108, pp. 96–114, 2017.
- [116] I. Naim, M. I. Tanveer, D. Gildea, M. E. Hoque, "Automated prediction and analysis of job interview performance: The role of what you say and how you say it," in *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, vol. 1, 2015, pp. 1–6, IEEE.
- [117] S. Peng, K. Nagao, "Automatic evaluation of presenters' discussion performance based on their heart rate," in *Proceedings of the 10th International Conference on Computer Supported Education - Volume 1: CSEDU*, 2018, pp. 27–34, INSTICC, SciTePress.

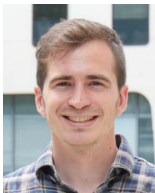
- [118] A. Shimada, Y. Taniguchi, F. Okubo, S. Konomi, H. Ogata, "Online change detection for monitoring individual student behavior via clickstream data on e-book system," in *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, LAK '18, New York, NY, USA, 2018, p. 446–450, Association for Computing Machinery.
- [119] F. J. Gutierrez, J. Simmonds, N. Hitschfeld, C. Casanova, C. Sotomayor, V. Peña Araya, "Assessing software development skills among k-6 learners in a project-based workshop with scratch," in *Proceedings of the 40th International Conference on Software Engineering: Software Engineering Education and Training*, ICSE-SEET '18, New York, NY, USA, 2018, p. 98–107, Association for Computing Machinery.
- [120] C. G. Brinton, M. Chiang, "MOOC performance prediction via clickstream data and social learning networks," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, 2015, pp. 2299–2307, IEEE.
- [121] M. Bouguessa, B. Dumoulin, S. Wang, "Identifying authoritative actors in question-answering forums: The case of Yahoo! Answers," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, New York, NY, USA, 2008, p. 866–874, Association for Computing Machinery.
- [122] W.-C. Kao, D.-R. Liu, S.-W. Wang, "Expert finding in question-answering websites: A novel hybrid approach," in *Proceedings of the 2010 ACM symposium on applied computing*, SAC '10, New York, NY, USA, 2010, p. 867–871, Association for Computing Machinery.
- [123] A. Borodin, G. O. Roberts, J. S. Rosenthal, P. Tsaparas, "Link analysis ranking: algorithms, theory, and experiments," *ACM Transactions on Internet Technology (TOIT)*, vol. 5, no. 1, pp. 231–297, 2005.
- [124] R. W. Herling, "Operational definitions of expertise and competence," *Advances in Developing Human Resources*, vol. 2, no. 1, pp. 8–21, 2000, doi: 10.1177/152342230000200103.
- [125] A. Santos, M. Souza, J. Oliveira, E. Figueiredo, "Mining software repositories to identify library experts," in *Proceedings of the VII Brazilian Symposium on Software Components, Architectures, and Reuse*, SBCARS '18, New York, NY, USA, 2018, p. 83–91, Association for Computing Machinery.
- [126] J. A. E. Montandon, L. L. Silva, M. T. Valente, "Identifying experts in software libraries and frameworks among GitHub users," in *Proceedings of the 16th International Conference on Mining Software Repositories*, MSR '19, 2019, p. 276–287, IEEE, IEEE Press.
- [127] C. Hauff, G. Gousios, "Matching GitHub developer profiles to job advertisements," in *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*, 2015, pp. 362–366.
- [128] M. S. Faisal, A. Daud, A. U. Akram, R. A. Abbasi, N. R. Aljohani, I. Mehmood, "Expert ranking techniques for online rated forums," *Computers in Human Behavior*, vol. 100, pp. 168 – 176, 2019, doi: <https://doi.org/10.1016/j.chb.2018.06.013>.
- [129] D. van Dijk, M. Tsagkias, M. de Rijke, "Early detection of topical expertise in community question answering," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, New York, NY, USA, 2015, p. 995–998, Association for Computing Machinery.
- [130] E. Constantinou, G. M. Kapitsaki, "Identifying developers' expertise in social coding platforms," in *2016 42nd Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 2016, pp. 63–67.
- [131] M. A. Vogel, *Leveraging information technology competencies and capabilities for a competitive advantage*. PhD dissertation, 2005.
- [132] D.-R. Liu, Y.-H. Chen, W.-C. Kao, H.-W. Wang, "Integrating expert profile, reputation and link analysis for expert finding in question-answering websites," *Information processing & management*, vol. 49, no. 1, pp. 312–329, 2013.
- [133] Z. Zhao, L. Zhang, X. He, W. Ng, "Expert finding for question answering via graph regularized matrix completion," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 4, pp. 993–1004, 2014.
- [134] A. Omidvar, M. Garakani, H. R. Safarpour, "Context based user ranking in forums for expert finding using wordnet dictionary and social network analysis," *Information Technology and Management*, vol. 15, no. 1, pp. 51–63, 2014.
- [135] A. Pal, F. M. Harper, J. A. Konstan, "Exploring question selection bias to identify experts and potential experts in community question answering," *ACM Transactions on Information Systems (TOIS)*, vol. 30, no. 2, pp. 1–28, 2012.
- [136] V. V. Vydiswaran, M. Reddy, "Identifying peer experts in online health forums," *BMC medical informatics and decision making*, vol. 19, no. 3, p. 68, 2019.
- [137] H. Zhu, E. Chen, H. Xiong, H. Cao, J. Tian, "Ranking user authority with relevant knowledge categories for expert finding," *World Wide Web*, vol. 17, no. 5, pp. 1081–1107, 2014.
- [138] A. Abdaoui, J. Azé, S. Bringay, N. Grabar, P. Poncelet, "Expertise in french health forums," *Health informatics journal*, vol. 25, no. 1, pp. 17–26, 2019.
- [139] C. G. Brinton, S. Buccapatnam, M. Chiang, H. V. Poor, "Mining MOOC clickstreams: On the relationship between learner behavior and performance," 2015.
- [140] A. Bozkurt, B. Aydin, A. Taşkıran, Koral Gümüüşoğlu, "Improving creative writing skills of EFL learners through microblogging," *The Online Journal of New Horizons in Education*, vol. 6, pp. 88–98, 06 2016.
- [141] T. Gonulal, "The use of Instagram as a mobile-assisted language learning tool," *Contemporary Educational Psychology*, vol. 10, pp. 309–323, 07 2019, doi: 10.30935/cet.590108.
- [142] R. Vesselinov, J. Grego, "Duolingo effectiveness study," *City University of New York, USA*, vol. 28, 2012.
- [143] Z. Liu, G. Xu, T. Liu, W. Fu, Y. Qi, W. Ding, Y. Song, C. Guo, C. Kong, S. Yang, et al., "Dolphin: A spoken language proficiency assessment system for elementary education," Apr 2020, pp. 2641–2647, ACM.
- [144] J. Dixon, C. Belnap, C. Albrecht, K. Lee, "The importance of soft skills," *Corporate Finance Review*, vol. 14, pp. 35–38, May 2010.
- [145] B. Cimatti, "Definition, development, assessment of soft skills and their role for the quality of organizations and enterprises," *International Journal for quality research*, vol. 10, no. 1, 2016.
- [146] T. Strobach, P. A. Frensch, T. Schubert, "Video game practice optimizes executive control skills in dual-task and task switching situations," *Acta Psychologica*, vol. 140, no. 1, pp. 13–24, 2012.
- [147] H.-S. Hsiao, C.-S. Chang, C.-Y. Lin, P.-M. Hu, "Development of children's creativity and manual skills within digital game-based learning environment," *Journal of Computer Assisted Learning*, vol. 30, no. 4, pp. 377–395, 2014.
- [148] D. R. Woods, A. N. Hrymak, R. R. Marshall, P. E. Wood, C. M. Crowe, T. W. Hoffman, J. D. Wright, P. A. Taylor, K. A. Woodhouse, C. K. Bouchard, "Developing problem solving skills: The McMaster problem solving program" *Journal of Engineering Education*, vol. 86, no. 2, pp. 75–91, 1997.
- [149] H.-C. Chu, C.-M. Hung, "Effects of the digital game-development approach on elementary school students' learning motivation, problem solving, and learning achievement," *International Journal of Distance Education Technologies (IJDET)*, vol. 13, no. 1, pp. 87–102, 2015.
- [150] M. Scriven, R. Paul, "Defining critical thinking. the critical thinking community: foundation for critical thinking," 2007.
- [151] A. Pal, R. Farzan, J. A. Konstan, R. E. Kraut, "Early detection of potential experts in question answering communities," in *International Conference on User Modeling, Adaptation, and Personalization*, 2011, pp. 231–242, Springer.
- [152] C. Unger, D. Murthy, A. Acker, I. Arora, A. Chang, "Examining the evolution of mobile social payments in Venmo," in *International Conference on Social Media and Society*, SMSociety'20, New York, NY, USA, 2020, p. 101–110, Association for Computing Machinery.
- [153] M. Yuan, L. Zhang, X.-Y. Li, H. Xiong, "Comprehensive and efficient data labeling via adaptive model scheduling," in *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, 2020, pp. 1858–1861, IEEE.
- [154] D. Papp, G. Szűcs, Z. Knoll, "Machine preparation for human labelling of hierarchical train sets by spectral clustering," in *2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 2019, pp. 157–162, IEEE.
- [155] J. Kang, S. Liu, M. Liu, *Tracking Students' Activities in Serious Games*, pp. 125–137. Cham: Springer International Publishing, 2017.
- [156] M. Oczak, K. Maschat, D. Berckmans, E. Vranken, J. Baumgartner, "Can an automated labelling method based on accelerometer data replace a human labeller?—postural profile of farrowing sows," *Computers and Electronics in Agriculture*, vol. 127, pp. 168–175, 2016.
- [157] C. A. Gomez-Urbe, N. Hunt, "The Netflix recommender system: Algorithms, business value, and innovation," *ACM Transactions on Management Information Systems (TMIS)*, vol. 6, no. 4, pp. 1–19, 2015.
- [158] A. Emerson, N. Henderson, J. Rowe, W. Min, S. Lee, J. Minogue, J. Lester, "Investigating visitor engagement in interactive science museum exhibits with multimodal Bayesian hierarchical models," in *Artificial Intelligence in Education*, Cham, 2020, pp. 165–176, Springer International Publishing.

- [159] A. K. Cooper, S. Coetzee, “On the ethics of using publicly-available data,” in *Responsible Design, Implementation and Use of Information and Communication Technology*, Cham, 2020, pp. 159–171, Springer International Publishing.
- [160] C. Kuhlman, L. Jackson, R. Chunara, “No computation without representation: Avoiding data and algorithm biases through diversity,” *arXiv preprint arXiv:2002.11836*, 2020.
- [161] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, “Certifying and removing disparate impact,” in *Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.
- [162] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations,” *Science*, vol. 366, no. 6464, pp. 447–453, 2019.



Sofia Strukova

She received the B.Sc. degree in computer science from Moscow Power Engineering Institute, Russia and M.Sc. degree in Big Data from the University of Murcia, Spain where she is currently pursuing the Ph.D. degree with the Department of Information and Communications Engineering.



José A. Ruipérez-Valiente

He received the B.Eng. degree in telecommunications from Universidad Católica de San Antonio de Murcia, and the M.Eng. degree in telecommunications and the M.Sc. and Ph.D. degrees in telematics from Universidad Carlos III of Madrid while conducting research with Institute IMDEA Networks in the area of learning analytics and educational data mining. He was a postdoctoral associate with MIT.

He has received more than 20 academic/research awards and fellowships, has published more than 90 scientific publications in high impact venues, and participated in over 20 funded projects and contracts. He is currently an Assistant Professor at the University of Murcia. More info at <https://webs.um.es/jruiperez>



Félix Gómez Mármol

He is a researcher in the Department of Information and Communications Engineering at the University of Murcia, Spain. His research interests include cybersecurity, internet of things, machine learning and bio-inspired algorithms. He received a M.Sc. and Ph.D. in computer engineering from the University of Murcia. More info at: <https://webs.um.es/felixgm>

Results of a Study to Improve the Spanish Version of the User Experience Questionnaire (UEQ)

Mónica Hernández-Campos^{1*}, Jörg Thomaschewski², Yuen C. Law¹

¹ Instituto Tecnológico de Costa Rica, Cartago (Costa Rica)

² University of Applied Sciences Emden/Leer, Emden (Germany)

Received 19 September 2021 | Accepted 26 October 2022 | Published 16 November 2022



ABSTRACT

This paper analyses changes in some items of the User Experience Questionnaire (UEQ) for use in the context of Costa Rican culture. Although a Spanish version of the UEQ was created in 2012, we use a double-translation and reconciliation model for detecting the more appropriate words for Costa Rican culture. These resulted in 7 new items that were added to the original Spanish version. In total, the resulting UEQ had 33 items. 161 participants took part in a study that examined both the original items and the new ones. Static analyses (Cronbach's Alpha, mean, variance, and confidence interval) were performed to measure the differences of the scales of the original items and the new UEQ variant with the Costa Rican words. Finally, confidence intervals of the individual items and Cronbach's Alpha coefficient average of the affected scales were analysed. The results show, contrary to initial expectations, that the Costa Rican word version is neither better nor worse than the original Spanish version. However, this shows that the UEQ is very robust to some changes in the items.

KEYWORDS

Evaluation, Questionnaires, User Experience.

DOI: 10.9781/ijimai.2022.11.003

I. INTRODUCTION

NOWADAYS, users expect devices, products and services that offer quite natural and easy-to-learn interactions. Especially the daily use of smartphones or tablets have brought the general expectation of users regarding the user experience of user interfaces to a high level, even if it is a complex business application. Simply said, users today expect a perfect user experience. A well-known definition of user experience is given in ISO 9241-210 (2019) [1]. Here, user experience is defined as “*user's perceptions and responses that result from the use and/or anticipated use of a system, product or service*” [1]. Thus, user experience is seen as a holistic concept that includes all types of emotional, cognitive, or physical reactions regarding the actual or even perceived use of a product or service that occur before, during, and after use. Still, the standard does not provide a clear list of factors or methods for measuring user experience.

In many cases, questionnaires are used to measure the user experience of products or services because UX questionnaires are easy to use, and a common quantitative way to measure user experience [2]. There are various UX questionnaires, such as meCUE [3], SUPR-Q [4], UEQ [5], [6], VisAWI [7], and Web-CLIC [8]. One goal of using a UX questionnaire is the idea of getting a better understanding of the own product or service and making appropriate improvements.

All steps in a testing process, including design, validation, adaptation, administration, and scoring, should be designed to minimize construct-irrelevant variance and promote valid score interpretations for all examinees in the intended population. Removing all barriers

allows for the comparable and valid interpretation of test scores for all examinees, which is central to the validity and comparability of test scores. For this reason, those responsible for all steps in the testing process should guarantee to minimize the potential threats to validation such as linguistic, communicative, cognitive, cultural, or age matters. These characteristics can impede some individuals in demonstrating their standing on intended constructs. Often, a product or service must be offered in different languages. Thus, the measurement of the user experience should also be carried out in the languages in which the tool is available, so that users can do the evaluation in their native language. One of the most critical aspects is language and its cultural variations. According to international standards in testing, it is necessary to avoid the use of language that has different meanings or connotations for the test-takers as well as the use of unfamiliar words [9].

The User Experience Questionnaire (UEQ) is the one of very few standard UX questionnaires available in many different languages. At the moment, 36 language versions are offered (see ueq-online.org). The language versions are usually conscientiously constructed and evaluated in the individual countries by local scientists. The UEQ maintainers then include the language version on their website ueq-online.org and often stay in touch with local language version scientists beyond that. The Spanish version of the UEQ has been carefully created and evaluated (see [10], [11]). Particularly in Latin America, regional variations of the language have developed in each country. Although the words in each variation are generally understood by native speakers in other countries, slight differences in usage and meaning might hinder communication. Differences between European Spanish and American Spanish are even greater.

In the case of the UEQ, unwanted and unknown effects of different meanings might exist for some of the items. Due to this

* Corresponding author.

E-mail address: mohernandez@itcr.ac.cr

large cultural Spanish language area and the related different use of words and meanings, requests for changes to the Spanish version of the UEQ are sent to the UEQ maintainers from different research groups, and for the case of Costa Rica, this is no exception. In order to ensure the fairness and validity of the test and to avoid a language bias, a new set of words are proposed for some of the items. There are two possible outcomes from this investigation: 1) The proposed new words are a better fit, in which case the results of the UEQ will better represent the users' experience; and 2) The UEQ is robust enough to accept modifications of some words, which will allow the use of words that are more familiar in the region, hence reducing the risk of item misinterpretation. Furthermore, there are also requests for adaptation of various items in other languages (e.g., for the French and Arabic versions), so the procedures and findings described here about Spanish adaptation are of more global importance.

This article analyses changes to the UEQ items to better understand the items in Costa Rica. For this purpose, 163 participants took part in a study that examined both the original items and the items with more culturally appropriate words.

II. CONSTRUCTION OF THE GERMAN VERSION AND SPANISH VERSION OF THE UEQ

The original German version of the UEQ was created by Laugwitz et al. in 2006 [5] using a data analytical approach. An initial item set of 229 potential items related to the concept of user experience was created in several brainstorming sessions with usability experts. This initial set was then reduced to an 80 items raw version of the questionnaire by an expert evaluation. These 80 items raw version was used in several studies. In these studies, 153 participants answered the 80 items. Finally, the scales and the items representing each scale were extracted from this data set by factor analysis (principal components, varimax rotation). Details concerning the construction process of the UEQ can be found in the works of Laugwitz and colleagues [5], [6].

The reliability (i.e. the scales are consistent) and validity (i.e. the scales really measure what they intend to measure) of the UEQ scales were investigated in 11 usability tests with a total number of 144 participants and an online survey with 722 participants. The results of these studies showed a sufficiently high reliability of the scales (measured by Cronbach's Alpha). As a result of this questionnaire construction, 6 scales with the following items were obtained.

Attractiveness: General impression towards the product. Do users like or dislike the product? The scale is a valence dimension. Items: *annoying/enjoyable, good/bad, unlikable/pleasing, unpleasant/pleasant, attractive/unattractive, friendly/unfriendly.*

Perspiciuity: Is it easy to understand how to use the product? Is it easy to get familiar with the product? Items: *not understandable/ understandable, easy to learn/difficult to learn, complicated/easy, clear/confusing.*

Efficiency: Is it possible to use the product fast and efficient? Does the user interface look organized? Items: *fast/slow, inefficient/ efficient, impractical/practical, organized/cluttered.*

Dependability: Does the user feel in control of the interaction? Is the interaction with the product secure and predicable? Items: *unpredictable/predictable, obstructive/supportive, secure/not secure, meets expectations/does not meet expectations.*

Stimulation: Is it interesting and exciting to use the product? Does the user feel motivated for a further use of the product? Items: *valuable/inferior, boring/exiting, not interesting/interesting, motivating/demotivating.*

Novelty: Is the design of the product innovative and creative? Does the product grab the user's attention? Items: *creative/dull, inventive/conventional, usual/leading edge, conservative/innovative.*

Attractiveness is a pure valence dimension and consists of 6 items. Perspicuity, Efficiency and Dependability measure the goal-directed aspects, while Stimulation and Novelty measure the non goal-directed aspects. These scales are each measured with 4 items (see list above). In total, there are 5 scales with 4 items each and the scale attractiveness with 6 items. The entire questionnaire thus consists of 26 items.

It is easy to see in the list above that each item of the UEQ consists of a pair of terms with opposite meanings. So, a semantic differential was chosen as item format, since this allows a fast and intuitive response. Each item can be rated on a 7-point Likert scale. Answers to an item therefore range from -3 (fully agree with negative term) to +3 (fully agree with positive term). Half of the items start with the positive term, the rest with the negative term (in randomized order).

Examples:

Not understandable o o o o o o *Understandable*

Efficient o o o o o o *Inefficient*

Applying the UEQ does not require much effort. Usually 3-5 minutes are sufficient for a participant to read the instructions and complete the questionnaire. The UEQ can either be used in a paper-pencil form or as an online questionnaire. Analysing the results of the UEQ is also no effort, as a comprehensive Excel tool is available for this purpose on the website. This Excel tool also contains a Benchmark [12] for a better interpretation of the result.

As described in Rauschenberger et al. [10], a Spanish version of the UEQ was created in 2012. First, the German version of the UEQ was translated into Spanish by two scientists with human computer interaction (HCI) and UEQ experience, a native Spanish speaker (living in Spain) and a bilingual scientist (native German, Spanish level C1, living in Germany). The translation was done in joint discussion for each item. During translation, the English version was also used to better align the items. Afterwards, the Spanish version was back-translated into German by an independent scientist (native German, Spanish level C2, living in Spain). If the words matched the original words, the translation was considered successful. Otherwise, the process was repeated until all words matched.

In a next step, the translation was checked with two different studies [11]. The web shop amazon.de and the communication software Skype were used, each with 94 participants. The two studies were conducted in Spain (Vigo) and found to have good internal consistency, determined with the Cronbach's Alpha [11].

Later, international comparative studies with different test objects have also confirmed the good appropriability of the results of the Spanish UEQ version and its internal consistency (e.g. [12]).

III. METHODS

As described above, the main purpose of this work was to adapt and validate the Spanish version of the original UEQ to Costa Rican culture. For this matter, we first translated the original German words to Costa Rican Spanish, using a double-translation and reconciliation model [14]. A native Costa Rican Spanish speaker with a C1 German level translated the words to Spanish, these were then translated back to German by a native German speaker who is familiar with Costa Rican Spanish (double-translation). From the resulting back-translated words, four pairs were completely different to the original German words. We reviewed and corrected the translations for these pairs (reconciliation). The resulting Spanish word list was finally compared to the Spanish version available at the UEQ website and the pairs that were completely different were selected. These resulted in 7 new items that were added to the original Spanish version. In total, the resulting UEQ had 33 items. The original items, the corresponding new items

and the affected scales are shown in Table I. The new UEQ was then applied in a study to compare the new items with their existing counterparts.

TABLE I. NEW AND ORIGINAL ITEMS WITH THE ITEM NUMBER IN THE QUESTIONNAIRE AND WITH THE RELATED SCALE IN PARENTHESES. FOR BETTER UNDERSTANDING, THE ENGLISH ITEMS ARE GIVEN IN THE LAST COLUMN

No	New items & (Scale)	No	Original Items & (Scale)	Engl. Items & (Scale)
27	Tedioso/ Ameno (Atracción)	1	Desagradable/ Agradable (Atracción)	Annoying/ Enjoyable (Attractiveness)
28	Incomprensible/ Comprensible (Transparencia)	2	No entendible/ Entendible (Transparencia)	Not understandable/ Understandable (Perspicuity)
29	Estorboso/ Facilitador (Controlabilidad)	11	Obstructivo/ Impulsor de apoyo (Controlabilidad)	Obstructive/ Supportive (Dependability)
30	Repugnante/ Llamativo (Atracción)	14	Repeler/ Atraer (Atracción)	Unlikable/ Pleasing (Attractiveness)
31	Molesto/ Placentero (Atracción)	16	Incómodo/ Cómodo (Atracción)	Unpleasant/ Pleasant (Attractiveness)
32	Según lo esperado/ Contrario a lo esperado (Controlabilidad)	19	Cubre expectativas/ No cubre expectativas (Controlabilidad)	Meets expectations/ Does not meet expectations (Dependability)
33	Poco práctico/ práctico (Eficiencia)	22	No pragmático/ Pragmático (Eficiencia)	Impractical/ Practical (Efficiency)

As described previously, the scale Attractiveness consists of 6 items and all other scales consist of 4 items. In Table I, it is seen that 3 items of the scale Attractiveness were modified (= 50%), 2 items of the scale Dependability were modified (= 50%), 1 item each of the scale Perspicuity (= 25%), and Efficiency (= 25%) were also modified. The items of the Stimulation and Novelty scales remained unchanged.

A. Procedure and Materials

The study was performed virtually, and all participants were asked to fill an online form. To start, participants read information and instructions about the study. This was followed by a short demographic questionnaire. Finally, they were presented with the 33 UEQ items.

Participants were explicitly asked to evaluate the “Netflix” application, but were also asked in the online form to write the name of the application they were evaluating. This was then used to validate the data (see Methods subsection).

B. Participants

163 participants were recruited during 2020 through the snowball strategy (56,4% male and 42,9% female). We shared the UEQ using social media asking for volunteers over 18 years old. They were not paid for their participation. All participants reported 100% experience using computers, and experience with the evaluated software “Netflix”.

C. Methods

From the 163 completed questionnaires, we filtered out those who wrote something different than “Netflix” in the corresponding field. The questionnaires of those participants who did not complete the

instrument seriously, for example, if all answers had the same value or if they were random, were also filtered out. For the latter case, we used a simple heuristic and checked the best and worst evaluations for the items in the same scale, if the difference was greater than 3 in any scale, all answers for that participant were discarded. In total, 2 questionnaires were excluded from this study, for a total of 161 valid questionnaires analysed.

To compare the new words to the original ones, we performed a series of tests first with the original set and then exchanging each of the seven items mentioned previously with its corresponding new word pair, for one-to-one comparisons, while for aggregated and average comparisons, all 7 items were substituted.

First, we compared the means, variance, standard deviation, and confidence intervals of the answers for the affected scales.

Following this, a Cronbach’s Alpha Coefficient was calculated in order to measure the consistency of the scales of the new UEQ variant with the Costa Rican words. This was compared to the consistency of the scales of the original UEQ. The user experience questionnaire contains 6 scales: attractiveness, perspicuity, efficiency, dependability, stimulation and novelty, but only 4 of these (attractiveness, perspicuity, efficiency, dependability) were affected by the new proposed items. Cronbach’s Alpha coefficient and confidence intervals were calculated for each of these scales according to Bonett [15].

Finally, sample sizes (precision, error probability) were used to compare both versions and factor analyses were carried out to find differences.

IV. RESULTS

The results are split into two parts. First, the mean values of the items and the scales of the Spanish original UEQ are compared with the Costa Rican UEQ. Then, a comparison of the Cronbach’s Alpha coefficients is made.

A. Results of the UEQ Comparing Mean Values

To compare the original UEQ and the new UEQ variant with the Costa Rican items, Table II shows the descriptive statistics of the original items compared to the new ones. Mean (M), standard deviation (SD) and variance (V) were calculated for the answers of the participants for each of these items. It can be seen in Table II that some mean values barely differed (Item No 28, 30, 31, 32), but other mean values lead to a noticeable difference (Item No 27, 29, 33). Thus, changes can be seen at the level of the individual items.

TABLE II. DESCRIPTIVE STATISTICS (MEAN, STANDARD DEVIATION AND VARIANCE): COMPARISON BETWEEN NEW AND ORIGINAL ITEMS

Item No	New			Item No	Original		
	M	SD	V		M	SD	V
27	1,5	1,2	1,3	1	2,4	0,5	0,7
28	2,1	0,7	0,9	2	2,4	0,6	0,8
29	1,8	1,1	1,2	11	0,6	1,3	1,6
30	1,9	0,9	0,7	14	1,9	1,0	1,0
31	1,9	0,9	0,9	16	2,1	0,9	0,8
32	1,2	1,5	2,3	19	1,2	1,5	2,1
33	1,9	1,2	1,4	22	1,3	1,1	1,2

An examination of the UEQ scales shows that the differences in the individual items can also result in different mean values in the overall result of the scales. Fig. 1 shows the mean values and the confidence interval with the changed items.

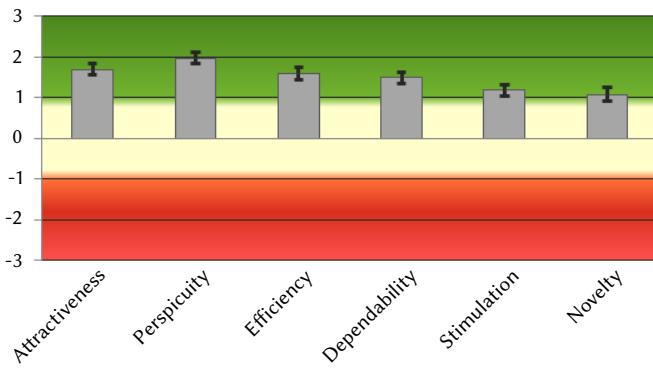


Fig. 1. Results of the evaluation of the test object “Netflix” by 161 subjects with the modified items with 5% confidence interval as error bar.

The mean values and the confidence interval of the original UEQ are shown in Fig.2.

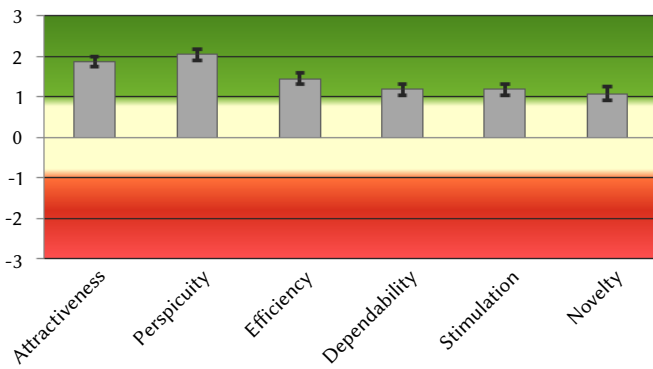


Fig. 2. Results of the evaluation of the test object “Netflix” by 161 test persons with the original Spanish version of the UEQ with 5% confidence interval as error bar.

For a more detailed comparison, the results of Fig. 1 and Fig. 2 have been combined in the Table III. Note that as previously described 3 items of the scale Attractiveness were modified (= 50%), 2 items of the scale Dependability were modified (= 50%), 1 item each of the scale Perspicuity (= 25%), and Efficiency (= 25%) were also modified. The items of the Stimulation and Novelty scales remained unchanged.

TABLE III. DESCRIPTIVE STATISTICS (MEAN, 5% CONFIDENCE): COMPARISON BETWEEN NEW AND ORIGINAL VERSION OF THE UEQ

Scale	New		Original	
	M	Conf	M	Conf
Attractiveness	1,69	0,13	1,87	0,12
Perspicuity	1,97	0,14	2,04	0,14
Efficiency	1,60	0,15	1,45	0,15
Dependability	1,49	0,14	1,18	0,13
Stimulation	1,18	0,14	1,18	0,14
Novelty	1,07	0,17	1,07	0,17

Further statistical results, in addition to the mean value, the standard deviation and the variance were also examined (see Table IV).

Since the same participants answered the questionnaire in both cases, a smaller variance can be interpreted as a better quality of a scale. According to this, the scales Attractiveness, Perspicuity, and Efficiency are better in the original version (see Table IV).

TABLE IV. DESCRIPTIVE STATISTICS (MEAN, STANDARD DEVIATION AND VARIANCE): COMPARISON BETWEEN NEW AND ORIGINAL VERSION OF THE UEQ

Scale	New			Original		
	M	SD	V	M	SD	V
Attractiveness	1,69	0,82	0,67	1,87	0,75	0,57
Perspicuity	1,97	0,93	0,86	2,04	0,88	0,78
Efficiency	1,60	0,99	0,98	1,45	0,95	0,90
Dependability	1,49	0,87	0,76	1,18	0,87	0,75
Stimulation	1,18	0,92	0,85	1,18	0,92	0,85
Novelty	1,07	1,09	1,20	1,07	1,09	1,20

By comparing the mean values, no statement can be made as to whether one of the two questionnaires is better suited to measuring “Netflix”. Therefore, a comparison of the Cronbach’s Alpha coefficients is made in the following section.

B. Comparison of the Cronbach’s Alpha Coefficients

The value of the Cronbach’s Alpha coefficient can be used as a degree of reliability. A significant higher value of the Cronbach’s Alpha coefficient can be a signal for an improvement of the UEQ scales.

In Table V, the average Cronbach’s Alpha coefficient is presented with 5% confidence interval for each scale.

TABLE V. AVERAGE CRONBACH’S ALPHAS AND CONFIDENCE INTERVALS FOR THE UEQ WITH NEW AND ORIGINAL ITEMS

Scale	New			Original		
	Avg. Alpha	Confidence interval		Avg. Alpha	Confidence interval	
Attractiveness	0,87	0,83	0,90	0,84	0,80	0,87
Perspicuity	0,73	0,65	0,79	0,66	0,56	0,73
Efficiency	0,66	0,57	0,74	0,63	0,52	0,71
Dependability	0,57	0,45	0,67	0,53	0,40	0,64
Stimulation	0,69	0,60	0,76	0,69	0,60	0,76
Novelty	0,72	0,64	0,78	0,72	0,64	0,78

Again, the same participants answered the questionnaire and thus a higher value for the Cronbach’s Alpha coefficient can be interpreted as better reliability of a scale. All Cronbach’s Alpha values have slightly improved in the new UEQ version, but are within the confidence interval of the values of the original UEQ version (see Table V).

The data of the Table V are shown as graph in Fig 3 too for easier comparison.

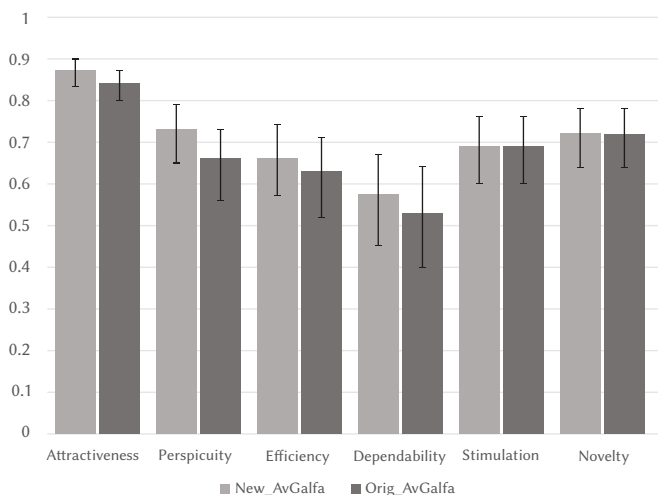


Fig. 3. Average Cronbach’s Alphas and confidence intervals (shown as error bars) for the UEQ with new and original items.

A general conclusion should not be made from the slight improvement of the Cronbach's Alpha coefficients (see Fig. 3), since on the one hand the increases are only small and on the other hand only the measurement of many different products could lead to a valid statement. These differences might be also be attributed to the fact that the new items were added at the end, which might have influenced the way in which the users responded, an in-between subjects test would be required to rule this out. Additionally, since the Cronbach's Alpha is quite sensitive in a scale with only 4 items, one might expect a significant change if 50% of the items are replaced. A good description of different effects with Cronbach's Alpha can be found in the work of Schrepp [16].

Another quality for the evaluation of questionnaire results is the Precision (deviation between true scale mean in the population and the estimated scale mean from the sample) and the Error-Probability, which can be calculated with the help of the standard distribution. These values can be taken from the Excel tool for the UEQ, as can all the values mentioned above (see www.ueq-online.org). In both cases, the new UEQ variant with the Costa Rican items and the original UEQ, have the same corresponding values for Precision=0.25 and Error-Probability=0.01 (related to N=161). Although this shows that the study with 161 participants led to a trustworthy result, an improvement through the new items cannot be read from this either.

Furthermore, factor analyses were carried out and the loading of the items to the factors was considered (un-rotated, promax rotation, varimax rotation). Here, too, no noteworthy difference between the new UEQ variant with the Costa Rican items and the original UEQ could be detected, which is mainly due to the fact that when only one test item is used (in this case "Netflix"), all items primarily load on one or at most two factors. This was also expected in advance and simply means that (almost) all items fit the test object. Only the measurement of many different products would provide a higher significance here.

The main result is: translated UEQ scales are very stable against deviations (replacement of individual items by items with at least very similar meaning).

V. CONCLUSIONS AND FURTHER WORK

In this study, items from the Spanish language version of the UEQ were adapted to the language culture in Costa Rica and evaluated in a study with 161 participants using the subject "Netflix". The aim of the study was to obtain an improved UEQ version for language use in Costa Rica. For this purpose, 7 item pairs were changed from the original UEQ and added to the original UEQ, so that the UEQ used in this study consisted of 33 item pairs.

Due to the widespread use of the UEQ in the Spanish-speaking community and the desire for a more culturally appropriate language version of the UEQ, this study is of great interest. But even beyond Spanish language differences, the results are interesting for all researchers and practitioners who would like to change individual items of the UEQ, as it provides the procedure to modify, add, and test new items.

We have demonstrated a procedure in which the items are not simply changed, but are appended to the original UEQ. In this way, a direct comparison is made with the same participants by conducting the evaluation with the original items on the one hand and with the changed items on the other. Thus, the effects of the changes can be directly compared with the results of the original UEQ.

In this study, we were able to show that changes to the items can lead to changed results. However, it is not possible to determine whether a modified questionnaire has a higher validity or reliability if only one product (here "Netflix") is evaluated.

It could be established that the UEQ behaves very robustly in the face of carefully implemented changes. Contrary to original expectations, the changes did not have as strong an effect as originally expected. This means that both the items of the original UEQ and the items in the new Costa Rica version were understood by the subjects. Thus, the new Costa Rican version, which uses words that are more familiar in the region, can also be used in further studies, reducing the risk of item misinterpretation by the users, although a cross-national comparison is then not possible.

It was also found that when only one product is evaluated, it is not possible to obtain statements about a clear improvement through statistical analyses. Additional studies are needed to get a clearer picture here.

In future studies, the translations of the newer UEQ+ [17] can be tested for this kind of robustness. The UEQ+ is a framework and currently provides 19 different scales, e.g. clarity [18]. From these 19 scales, a questionnaire is created that fits the product [19].

REFERENCES

- [1] ISO 9241-210, "Ergonomics of human-system interaction - part 210: Human-centred design for interactive systems," International Organization for Standardization, Geneva, Switzerland, 2019.
- [2] J. Lazar, J.H. Feng, and H. Hochheiser, *Research methods in human-computer interaction*, 2nd ed., Glasgow, United Kingdom: Bell & Brain, 2010.
- [3] M. Minge, and L. Riedel, "meCUE - Ein modularer Fragebogen zur Erfassung des Nutzererlebens [meCue - A modular questionnaire for capturing the user experience]," in S. Boll, S. Maaß and R. Malaka (Ed.): *Mensch und Computer [Humans and computers] 2013: Interaktive Vielfalt [Interactive diversity]*, Oldenbourg Verlag, München, pp. 89-98, 2013.
- [4] J. Sauro, "SUPR-Q: A comprehensive measure of the quality of the website user experience," *Journal of Usability Studies*, vol. 10, no. 2, pp. 68-86, 2015, doi: 10.5555/2817315.2817317.
- [5] B. Laugwitz, M. Schrepp, and T. Held, "Konstruktion eines Fragebogens zur Messung der User Experience von Softwareprodukten [Construction of a questionnaire for the measurement of user experience of software products]," in A.M. Heinecke and H. Paul (Eds.): *Mensch & Computer [Humans & computers] 2006*, Oldenbourg Verlag, pp. 125 - 134, 2006.
- [6] B. Laugwitz, T. Held, and M. Schrepp, "Construction and evaluation of a user experience questionnaire," in *Symposium of the Austrian HCI and Usability Engineering Group*, Heidelberg, Germany, pp. 63-76, 2008, doi: 10.1007/978-3-540-89350-9_6.
- [7] M. Thielsch and M. Mooshgen, "Erfassung visueller Ästhetik mit dem VISAWI," *Usability Professionals 2011*, Stuttgart, Germany, pp. 260-265, 2011.
- [8] M. Thielsch and G. Hirschfeld, "Facets of website content," *Human-Computer Interaction*, vol. 34, no.4, pp. 279-327, 2019.
- [9] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, "Standards for educational and psychological testing," *American Educational Research Association*, 1999.
- [10] M. Rauschenberger, M. Schrepp, S. Olschner, J. Thomaschewski, and MP. Cota, "Measurement of user experience: A Spanish Language Version of the User Experience Questionnaire (UEQ)," In: Á.J.A. Rocha, L.P.R. Calvo-Manzano, M. Pérez Cota (editors), *Information Systems and Technologies (CISTI)*, Madrid, Spain, pp. 471-476, 2012.
- [11] M. Rauschenberger, M. Schrepp, MP. Cota, S. Olschner, and J. Thomaschewski, "Efficient measurement of the user experience of interactive products - How to use the User Experience Questionnaire (UEQ). Example: Spanish Language Version," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 2, no. 1, pp. 39-45, 2013, doi: 10.9781/ijimai.2013.215.
- [12] M. Schrepp, A. Hinderks, and J. Thomaschewski, "Construction of a Benchmark for the User Experience Questionnaire (UEQ)," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 4, pp. 40-44, 2017, doi: 10.9781/ijimai.2017.445.
- [13] A. Hinderks, M. Schrepp, F. J. Domínguez Mayo, M. J. Escalona, and J. Thomaschewski, "Developing a UX KPI based on the user experience questionnaire," *Computer Standards & Interfaces*, no. 65, pp. 38-44, 2019,

doi: 10.1016/j.csi.2019.01.007.

- [14] International Test Commission, “The ITC Guidelines for Translating and Adapting Tests (Second edition),” 2017. Accessed: June 14 2021. [Online]. Available: www.intestcom.org/files/guideline_test_adaptation_2ed.pdf
- [15] D.B. Bonett, “Sample Size Requirements for Testing and Estimating Coefficient Alpha,” *Journal of Educational and Behavioral Statistics*, vol. 27, no. 4, pp. 335-340, 2002, doi: 10.3102/10769986027004335.
- [16] M. Schrepp, “On the Usage of Cronbach’s Alpha to Measure Reliability of UX Scales,” *Journal of Usability Studies*, vol. 15, no. 4, pp. 247–258, 2020.
- [17] M. Schrepp, and 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, pp. 88-95, doi: 10.9781/ijimai.2019.06.006.
- [18] M. Schrepp, R. Otten, K. Blum, and J. Thomaschewski, “What Causes the Dependency between Perceived Aesthetics and Perceived Usability?,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 78-85, 2021, doi: 10.9781/ijimai.2020.12.005.
- [19] A.-L. Meiners, J. Kollmorgen, M. Schrepp, and J. Thomaschewski, “Which UX Aspects Are Important for a Software Product?,” In: S. Schneegass, B. Pflöging, and D. Kern (editors): *Mensch und Computer (MuC)*, Ingolstadt, Germany, pp. 136–139, 2021.



Mónica Hernández-Campos

MSc. Mónica Hernández-Campos obtained her Master’s Degree in Cognitive Sciences at the University of Costa Rica. She is an active student of the Doctorate “Formación en la Sociedad del Conocimiento” of the University of Salamanca. She is a psychology teacher and academic advisor at the Costa Rica Institute of Technology. Her research interests are cognition, learning, and educational innovation.



Jörg Thomaschewski

Dr. Jörg Thomaschewski received a PhD in physics from the University of Bremen (Germany) in 1996. He became Full Professor at the University of Applied Sciences Emden/Leer (Germany) in September 2000. His research interests are human-computer interaction, e-learning, and software engineering. Dr. Thomaschewski is founder of the research group “Agile Software Development and User Experience”.



Yuen C. Law

Dr. Yuen C. Law obtained his PhD in computer Science from the RWTH Aachen University in 2016 and currently works at the Computer Science department at the Costa Rica Institute of Technology as docent and researcher. His interests include virtual and augmented reality applications, visualization, and human-computer interaction.

Tests of Usability Guidelines About Response to User Actions. Importance, Compliance, and Application of the Guidelines

Lucía Alonso-Virgós¹, Jordán Pascual Espada², Gustavo Rossi³ *

¹ Universidad Internacional de La Rioja (Spain)

² Universidad de Oviedo (Spain)

³ Facultad de Informática, UNLP and Facultad de Tecnología Informática, UAI (Argentina)

Received 5 January 2021 | Accepted 4 January 2023 | Published 29 November 2023



ABSTRACT

Usability is a quality that a web page can have due to its simple use. Many recommendations aim to improve the web user experience, but there is no standardization of them. This study is part of a saga, which aims to order existing recommendations and guidelines by analyzing the behavior of 20 Information Technology (IT) developers. This publication analyzes the set of guidelines that determine "user responses" when they interact with a website. It is intended to group these guidelines and obtain data on the application of each of them. The test is carried out with 20 web developers without training or experience in web usability. The objective is to know if there are "user response" guidelines that a developer with no training or usability experience applies innately. Since web developers are also users, it is believed that there may be innate behavior that is not necessarily learned. The purposes of the work are: 1) Enumerate the most forgotten recommendations by web developers. This can help to think about the importance of offering specific training in this field. 2) Know the most important recommendations and guidelines, according to the web developers themselves. The investigation is carried out as follows: First, IT engineers were asked to develop a website; Second, user tests were performed and the most neglected and most applied guidelines were evaluated. The level of compliance was also analyzed, as developers lack experience in web usability and could be applying a guideline, but not correctly; Third, web developers are interviewed to find out what guidelines they consider necessary. The results are intended to help us understand if a web developer without training or experience in web usability can innately apply guidelines on "user responses". The objective of the study is to determine that there are guidelines that are applied intuitively and others that are not, and to know the reason for each situation. The results determine that the guidelines considered essential and those that are most applied innately have something in common. The results reveal that the essential guidelines and those that are most commonly implemented inherently share certain commonalities.

KEYWORDS

Action, Guidelines, Recommendations, Usability, User Response, User Experience, Web.

DOI: 10.9781/ijimai.2023.11.003

I. INTRODUCTION

WEB usability measures the quality of a user's experience when interacting with a web page. To measure the experience, the relationship between the website and who uses it is analyzed. A web page, or website, refers to the navigation system, its contents, and the functionality it offers.

Thus, web usability aims to facilitate a user to use a website efficiently. This efficiency involves the access of the elements offered on the screen and the fulfillment of the tasks that the user

intends. Many suggestions are published that improve the usability of web portals [1]. These "ideas" are classified into recommendations, heuristics, guidelines, etc. [2]. All these concepts are different and, therefore, seek different objectives.

Heuristics are design principles that allow interaction to be facilitated. The most popular ones were published by Jakob Nielsen in his book 10 Heuristics of Usability for User Interface Design (1995) [3]. They are useful, but experts have shown that their approach, mainly theoretical, is not the best answer to specific problems [4].

The guidelines have a similar objective to heuristics [5], [6]. Their foundations do not offer a theoretical framework that is broad enough to determine generality and applying them is more effective in specific cases [6]. So they are not always the best option because they are still too theoretical.

In our previous research, we have proposed to establish usability standards. Usability recommendations are the most useful for this [7].

* Corresponding author.

E-mail addresses: lucia.alonso.virgos@unir.net (L. Alonso Virgós), jordansoy@gmail.com (J. Pascual Espada), gustavo@lifa.info.unlp.edu.ar (G. Rossi).

However, and although many lists of recommendations have been published [7], to date they have not been grouped, classified, or sorted in a standardized manner. Getting recommendations to be grouped, classified, and ordered would be very useful for web developers. This is one of the objectives of our research.

For our research, 103 recommendations were extracted from different sources. Within this selection, usability recommendations for specific domains have been avoided [8]. Next, the 103 recommendations are divided into five groups that offer a classification [8]. Classifying recommendations helps in avoiding repetitions. The proposed groups are:

- (1) Recommendations to reduce “noise”
- (2) Follow conventions
- (3) Provide information quickly and understandably
- (4) Efficient and understandable controls for users to enter information
- (5) Give descriptive and understandable responses to user actions

After designing this ordered classification of usability recommendations [8], and after evaluating the recommendations of groups (1), (2), and (4), three scientific articles that evaluate these groups of recommendations [8], [9] were published.

This paper aims to evaluate the group (5) “Give descriptive and understandable responses to user actions”.

Of the 103 recommendations, there are 4 useful usability recommendations on this group [10]-[20].

It should be noted that the tests and interviews that were conducted during this investigation, involved participants without training or experience in web usability.

The participants are web developers. The idea of asking that these participants be newbies in web usability aims to make the evaluation objective and to be able to measure innate behavior during web development.

The mechanics of our research has been the following.

In addition to offering the above 103 grouped recommendations, participants are selected for tests and interviews. The participants are 20 web developers without training or experience in web usability.

1. Each participant develops a web portal referred to a specific objective, each one chooses their own.
2. Specific training on web usability is offered. This training is also divided into 5 blocks, so the knowledge of each guideline is acquired with precision.
3. When the participants have already received the appropriate training, their web developments are evaluated. This evaluation is made in 5 parts, coinciding with the groups. The application of each of the corresponding group guidelines and their level of compliance is measured.
4. A list of the groups is offered to the participants. They are interviewed to analyze the importance they attach to each of the guidelines. There are 5 interviews, one for each group. In this way, the results are more accurate.
5. Conclusions are drawn in this regard.

Though web developers have enough skills to develop websites, the purpose of this research is to assess if these skills (together with intuition) are sufficient to create usable Web sites and to measure objectively the deviation from this objective.

That is, what we intend to know is if a web developer intuitively applies web usability recommendations. And if compliance with the recommendation is correct. Or if, on the contrary, an IT engineer needs specific training on web usability, in addition to the acquisition of web development skills.

The research is divided into two objectives:

- Objective 1 is intended to determine the degree of application and the level of compliance with each of the fifth Group’s guidelines “Give descriptive and understandable responses to user actions”, by IT engineers with no training or experience in web usability.
- Objective 2 aims to know the importance that web developers give to each recommendation, after understanding its purpose. That is, after receiving training in web usability.

This article is organized as follows. In Section II we present some background on usability evaluation. In Section III heuristics and recommendations are extracted and grouped. In Section IV the research design is described. In Section V, results for each of the recommendations are presented and in Section VI we evaluate these results. In Section VII we discuss these results presenting the best and worst recommendations as valued by volunteers. In Section VIII we present our conclusions and suggest some topics for further research.

II. BACKGROUND

A. Usability Evaluation

Web usability not only measures the ease with which you browse a Web site, but also the effectiveness and efficiency with which it is done [10]. Many methods measure this.

There are three broad ways to measure usability:

(i) Usability inspections. These inspections are abstract concepts that are supported by expert studies or observations. The most common are heuristic evaluations, cognitive patterns, and checklists [6], [11], [12]. Their purpose is to evaluate specific actions or problems [13], [14], [15].

(ii) User-centered methods. Unlike the previous ones (i), users participate in these tests. This means that they are more practical than theoretical. They are tests, physiological measurements, or interviews [16], [17], [18]. Web usability is measured by looking for potential problems, mainly. Through the interviews, you can know the opinions of the users. In this case, several questions are asked about their behavior, attitude, thoughts, and feelings during web browsing. Through physiological measurements or monitoring, physiological responses are obtained from users to a website. For example, a physiological measurement is to use eye-tracking to measure the movement of the retina and know which areas of the interface are the ones that stand out, and which ones go unnoticed. Through the tests, it is possible to measure the efficiency in the interaction of the user with the computer. For example, you can measure the memory capacity that a user has while browsing the web on a website that he had not visited for a while. It is also possible to assess satisfaction by analyzing the facial expressions of users [19].

(iii) There is a third method of website evaluation: an automatic evaluation. Since experts in the field of web usability recommend that measurements be made by people, in this work this method is not considered. This statement is justified because the evaluation aims to discover the ease of use of a website. And the ease of use comes from the intuition of the user (as a person). The great advantage of automatism is that they are objective in their evaluation. However, if the evaluation is carried out by two people, they may offer different results due to subjectivity [32]-[35]. For research, we start from the fact that evaluations must be carried out by a person, and not by automatism [20].

The proposals of this project are evaluated through interviews answered by web developers, without experience in usability, who also think like users, and through expert analysis.

B. Heuristics, Guidelines, and Recommendations

As stated earlier, the heuristic method aims to discover and improve human-computer interaction. Nielsen [21], [22] stands out in this field, although other proposals designed for specific domains [23], [24] [25], [26] are also useful. The intention of these proposal is to detect existing problems and create a theoretical action plan that avoids errors [27]. As they are planned for specific domains, they are not useful for evaluating general web design problems [4].

An example of heuristics would be: “This (concrete) website must differentiate text links” or “Single-column paragraphs are read faster than multiple column paragraphs” [6], [28]. The solutions are given for specific websites and may or may not be useful for other websites. Besides, there is no standard to follow [6], [29], [30].

It has already been mentioned that this research aims to select the most important general guidelines. Bibliography used to extract these guidelines [6], [11], [27], [28], [30], [31], [32], [33] includes 103 generic recommendations [8], [28], [33] that are useful for any domain. These recommendations were extracted and analyzed for the purpose of this study.

An IT engineer is technically trained to develop a website but does not always seem trained in human-computer interactions. When it is not, its developments may not meet the needs of users [34] or even the specific needs of a domain [35], [36]. This “ignorance” or lack of training in human-computer interactions causes the application of web usability guidelines to be useless [37]. For all this, our project aims to discover if there is intuition during the application of the usability guidelines and if this application is fulfilled correctly. It is intended to demonstrate that there are recommendations for web usability that are intuitively applied and others that are not used, and the reason for each situation.

III. RECOMMENDATIONS

This publication focuses on the recommendations group (5) “Give answers descriptive and understandable to the actions of the users”. We aim to analyze these four useful recommendations with the assistance of 20 graduate students specializing in web engineering. The purpose is to know if these recommendations are applied innately and compliance is correct without the need for training. We also discover the importance that these IT engineers give to each recommendation once they understand their purpose.

A. Classification of Recommendations

The 103 recommendations were extracted from different sources [6], [11], [42], [43] and divided into groups by our teammate Jordán Espada, who analyzed the 103 recommendations and their objectives and looked for similarities to be able to group them. He found five viable similarities, differentiated by their purpose, and created the 5 groups.

- (1) Recommendations to reduce “noise”
- (2) Follow conventions
- (3) Give information quickly and comprehensibly
- (4) Efficient and understandable controls for users to enter information
- (5) Give answers descriptive and understandable to the actions of the users

This grouping proposal has been designed and serves as didactic material in the Master in Web Engineering of the University of Oviedo.

Fig. 1 shows the grouping of web usability recommendations [38]. The 103 recommendations taken from different sources are ordered and reduced to 69. As indicated in Fig. 1, repeated or overly specific recommendations are eliminated. They are divided into 5 groups, of 16, 8, 24, 17, and 4 recommendations.

Given that groups 1-4 have already been published, this article presents the recommendations from Group (5) - Provide clear and understandable responses to user actions.

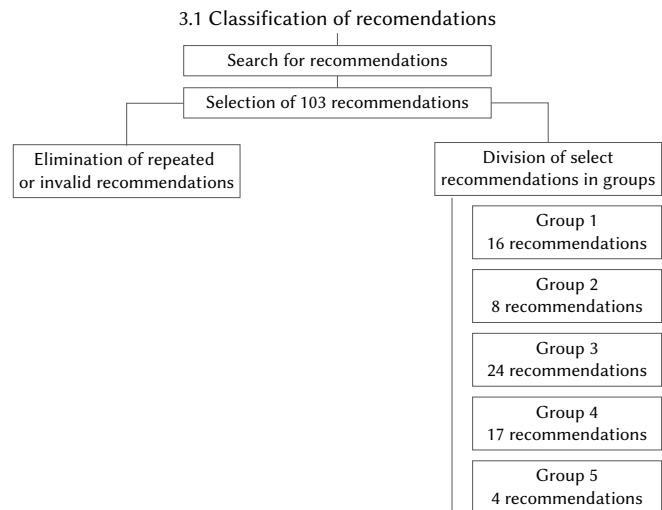


Fig. 1. Classification of recommendations.

1. Recommendation A: Being Able to Easily Identify Items Seen or Visited

This recommendation focuses on the website recognizing those elements that were visited or selected.

For example, in a specific search of the web browser the websites in which the user has previously entered are presented in purple. The rest of the websites that you have not visited yet appear in blue. The elements that must be recognized are those on which the user applied actions of importance [39].

In Fig. 2 the second search result is represented in purple because it has been visited before. In this way, it alerts the user that this site is already “read”.

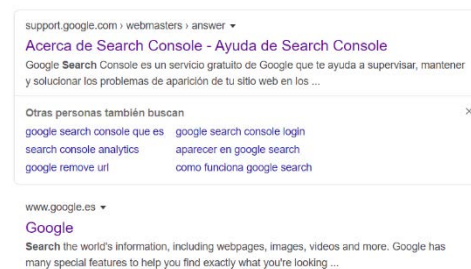


Fig. 2. Identify viewed or selected items. Google.

Fig. 3 identifies unread emails (with white background), emails already read (gray background). This is very useful for the user to quickly locate those emails that are unopened.

2. Recommendation B: Notice of Response to Actions

This recommendation intends that all user actions have outstanding notifications. For example, using color codes and standardized icons to represent the type of notification. These notifications can be errors, successes, warnings, etc. [40]-[41].

In Fig. 4, a notification is presented for a product that has been added to the shopping basket. Alongside the notification, additional options are available, such as editing the basket or proceeding to checkout. However, the importance of this recommendation is underscored by the clear notification of a new item being added to the basket.

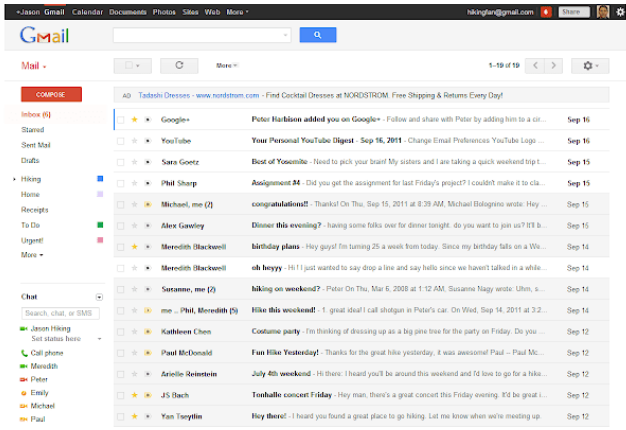


Fig. 3. Identify viewed or selected items. Gmail <https://chrome.google.com/webstore/detail/gmail/pjkljhegncpnkpnkbncohdijeoejaedia?hl=es-419>.

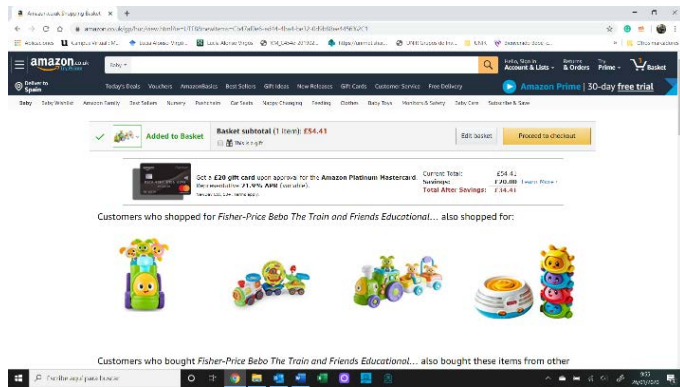


Fig. 4. Identify added to basket. Amazon.

A notification is sent in Fig. 5 that warns that a conversation has just been deleted to the recycle bin. With this notice, the user can be satisfied knowing that he or she has deleted the conversation if that was his or her intention, or the user can rectify (Recommendation C) if the deletion of the conversation is a mistake.

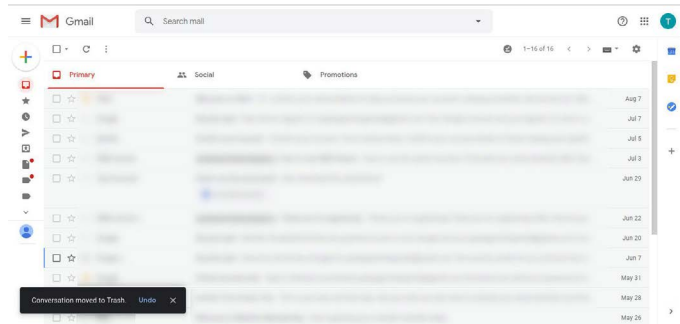


Fig. 5. Notice of response to actions. Gmail.

3. Recommendation C: “Undo” to Go Back on Tasks Sensitive

This recommendation allows user error tolerance. For example, requiring confirmation on some important tasks, such as a purchase. Although, sometimes, it is more efficient to use Undo than to request confirmation [42].

The purpose is to ensure that the user did not act by mistake.

As shown in Fig. 5, in addition to notifying the action that the user has just performed (delete a conversation), the user can go back by clicking on the “Undo” text link.

4. Recommendation D: Descriptive Information About Errors

This recommendation is intended to provide detailed information about an error, for example, why it occurred.

Sometimes you need to know if the error was made by the user or the system, or if there are problems with the information entered (for example, if unsolicited content has been sent as an unsupported symbol, if expected content is absent, if there is blank or invalid size/format content, or if there is content not validated by business logic). Fig. 6 shows an example of descriptive information about errors in Dropbox application, when entering an email in use [43].

When this happens, it should be clear:

- What error has occurred
- Where has it happened?
- How can it be solved

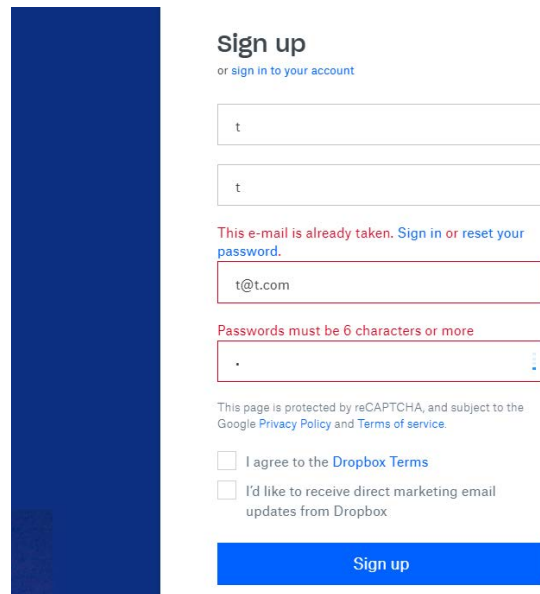


Fig. 6. Sing up in Dropbox. Dropbox.

IV. RESEARCH DESIGN

This project studies the behavior of 20 Spanish computer engineers. The purpose is to discover if a web developer with no experience in web usability applies a useful recommendation intuitively because the intuition factor is relevant. If the existing recommendations are not applied intuitively, it can be deduced that IT engineers need to be trained in web usability.

Of the 20 students, 15 are men and 5 are women and have an average age of 23 years. They are students of the Master in Computer Engineering at the University of Oviedo. Everyone has a degree in Web Engineering, so everyone has the technical capacity to develop a website. However, none have experience in web usability. Web usability is a block of didactic content that will be taught in the Master in Computer Engineering at the University of Oviedo after the experiment described in this article.

As no participant formally knows web usability, despite having a high level of knowledge in web development, our team wants to know if an IT developer of these characteristics can innately apply any of the recommendations of web usability. This hypothesis is based on the basis that states that web usability is easily detectable by a user, and IT developers are users in addition to web developers.

It is also intended to measure the level of compliance with the recommendations that have been applied. Besides, finally, it is expected to know the importance that the web developers themselves give to each recommendation after training in usability.

For the experiment, a subject is assigned to each student (banking, restaurants, etc.). From this topic, they should design a website.

Therefore, 20 different websites are created by IT engineers who ignore web usability. After receiving training, participants are trained to evaluate their web designs. In this particular case, students receive training on the 4 recommendations collected and grouped in Group (5).

The purpose is to know the importance that a web developer recently trained in web usability attributes to each of the group's recommendations, and if any of the developers apply the usability recommendations innately and correctly (see Fig. 7).

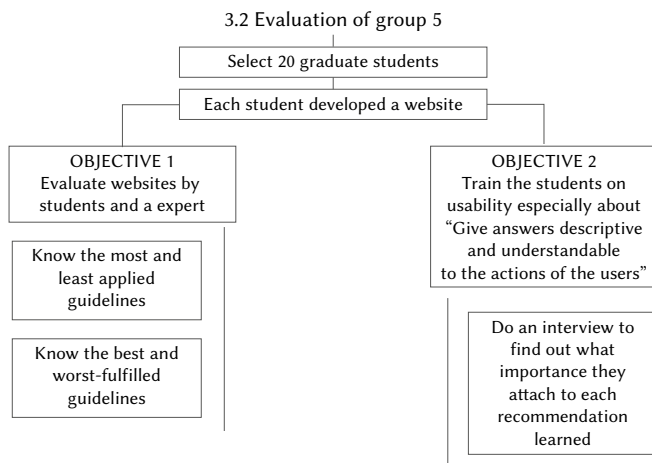


Fig. 7. Evaluation of group 5. Objective 1 and Objective 2.

To analyze the importance that each IT gives to each recommendation of the Group (5), surveys are carried out, (see Fig. 7, left side). The level of importance is measured with a score of 0-10 (Objective 1). As indicated above, the survey is conducted after web usability training to ensure that participants respond with knowledge. The measurement results are shown in Section V.

Next, each website is tested. The tests are carried out by the same participants, but this time assisted by an expert supervisor in web usability.

Website measurements are scored always following the same criteria. An applied recommendation is scored with 1, and an unapplied recommendation is scored with 0. The application measures the participant's intention to develop based on web usability. These results can provide useful information on whether web usability recommendations are innately used by inexperienced developers.

The degree of compliance is also measured with values from 0 to 5. 0 means that the recommendation is not properly fulfilled. 5 means that the recommendation is met successfully. It may be the case where a recommendation has a value of 1 in application and 0 in compliance. This means that the recommendation is applied innately because it is considered useful even without notions in web usability, but it is applied incorrectly (Objective 2). The results of the tests performed can be found in Section V of this article.

The summary of the process is:

1. Postgraduate students, already experienced web developers, were tasked with creating websites on assigned topics.
2. The students underwent training in usability, specifically focusing on the guidelines presented in this document (OBJ 1).

3. A survey was conducted to gather the students' perceptions regarding the application, fulfillment, and importance of the guidelines after the usability training (OBJ 2).

4. Each website was assessed by a usability expert to determine which guidelines were applied intuitively, which were not, and how this relates to the opinions expressed by each participant.

This comprehensive approach aims to understand the intuitive application of usability guidelines by inexperienced developers and how this aligns with their perceptions and usability training.

V. RESULTS OF THE EXPERIMENT

This section includes the results of tests a) and b) both with the 4 recommendations previously seen.

A. OBJECTIVE 1. Test A) "Innate" Use of Usability Guidelines by Developers

The first part of the experiment consists in that web developers create a website. After the development task, they are trained in web usability, particularly in the Group's recommendations (5). Once web developers have been trained in web usability, they attach importance to each of the guidelines, according to their criteria. With well-established knowledge, they evaluate their websites and measure the degree of application and the level of compliance in each of the 4 guidelines of this group. This step is taken with the help of an expert.

The application of a guideline is scored with a 1. The non-application of a guideline is scored with a 0. This non-application means that the guideline should have been used but was not used. On some occasions, the guidelines were used, value 1, but were not met properly. For this reason, compliance is studied in the following section of the paper.

Fig. 8 indicates that most applied guidelines belong to type B (Notice of response to actions) and D (Descriptive information about errors). 90% of the participants applied them. The least applied guideline is A (Identify viewed or selected items). In this case, 60% of web developers applied it. Guideline C ("Undo" to return to sensitive tasks) was applied by 75% of IT developers.

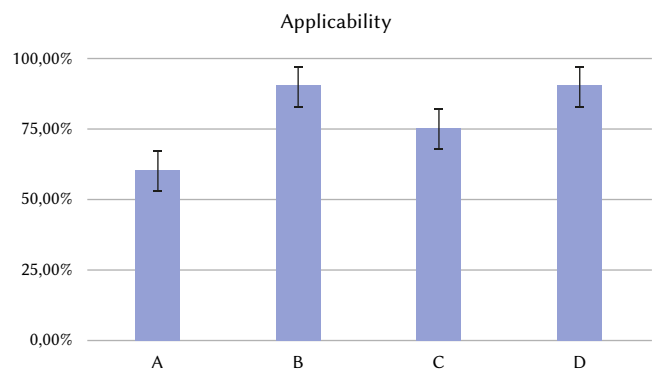


Fig. 8. Results of applicability.

Fig. 9 indicates that the major guideline complied with is D (Descriptive information about errors), with 78.5% compliance; then, the B (Notice of response to actions), with 77% compliance; these guidelines are applied in 90% of cases and three-fourths of the time were correctly applied. The worst compliments are the A (Identify viewed or selected items) and the C ("Undo" to go back on sensitive tasks), with 12% compliance. Guideline A stands out for properly fulfilling only 6% of the occasions. This recommendation is applied more than half of the time (60% of the time) and is incorrectly applied almost always.

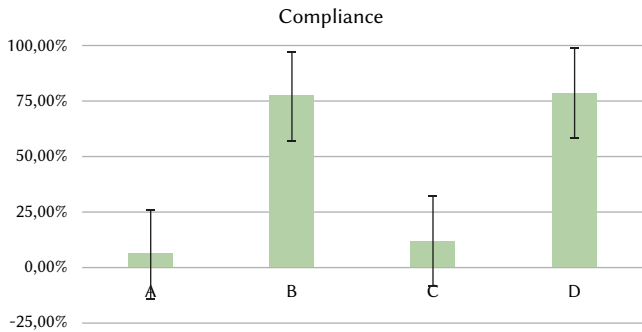


Fig. 9. Results of compliance.

B. OBJECTIVE 2. Test B) "Importance" Use of Usability Guidelines by Developers

Fig. 10 represents the degree of importance that web developers attach to each recommendation. We have already published three groups before this, and we know that the levels of importance granted are usually high. In the case of this experiment, the recommendation considered the most important is C ("Undo" to go back on tasks sensitive), with 85%. Then the A (Identify viewed or selected items) with 82.5%. It is followed by recommendation B (Notice of response to actions) with 81%. Finally, the recommendation considered less important is the D (Descriptive information about errors) with 67%. Curiously, recommendation A has been considered one of the most important, having received training in web usability, and yet only 60% of IT developers have applied it (in an innate form). Besides, this application has been quite wrong, with only 6% compliance. A similar case occurs with recommendation C, the most important with 85% and, although 75% of IT developers have applied it (in an innate way), only 12% have complied properly.

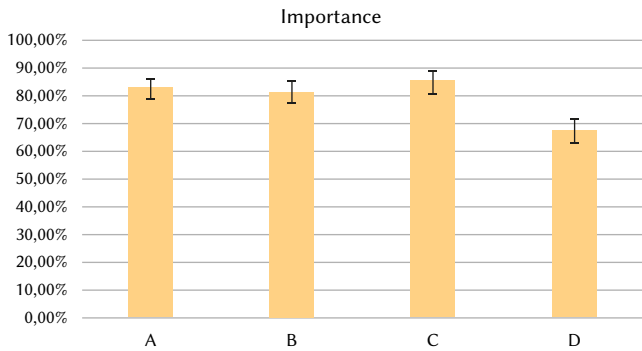


Fig. 10. Results of importance.

Fig. 11 compares the degree of importance given to each recommendation and the result of the Pearson coefficient of applicability.

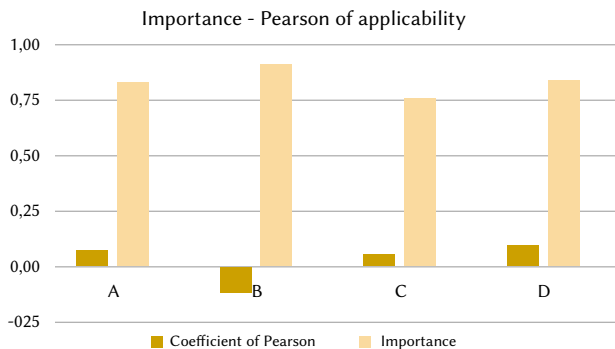


Fig. 11. Statistical analysis of the importance - Pearson of applicability.

Our research only publishes the scatter plots of those recommendations whose relationship is moderate, with a value $-0.40 >$ or < 0.40 , or high, with a value $-0.60 >$ or < 0.60 . In this case, there is no objective relationship in any of the recommendations.

Fig. 9 establishes that the recommendations closest to a relationship of variables are the B: (Notice of response to actions) and the D: (Descriptive information about errors). But in none, there is a considerable relationship. That is, it can be concluded that all recommendations seem to contradict the theory of relationships, with the variables studied in this analysis of Importance.

The objective is to determine if a higher Importance causes a higher Applicability rate. That is if the guideline that is most applied is also considered the most important and vice versa.

Unlike Covariance, Pearson's Correlation is independent of the scale of measurement of the variables. We use Pearson because the study aims to obtain the same index in all recommendations.

To interpret the results in detail, the dispersion diagram should be consulted. This diagram is used to analyze the strength and direction of the relationship between the variables. Although there have been no relationships in this analysis, the process is explained. The value of the correlation coefficient can vary from -1 to $+1$. The higher the absolute value of the coefficient, the stronger the relationship between the variables. An absolute value of 1 indicates a perfect linear relationship. A correlation close to 0 indicates that there is no linear relationship between the variables. In the analysis, the ratios obtained are: 0.07, -0.12 , 0.05 and 0.10, values too low to interpret that there is a relationship.

The sign of the coefficient indicates the direction of the relationship. If both variables tend to increase or decrease at the same time, the coefficient is positive and the line representing the correlation forms an upward slope. This occurs with guidelines A: (Identify viewed or selected items), C: ("Undo" to go back on tasks, and D: (Descriptive information about errors), whose results are positive (> 0).

If one variable tends to increase while the other decreases, the coefficient is negative and the line representing the correlation forms a downward slope. This happens with guideline B: (Notice of response to actions), whose result is negative (< 0).

C. Analysis of the IMPORTANCE - Pearson of COMPLIANCE

The relationship pattern between the Pearson coefficient of Compliance and Importance variables is analyzed. Conclusions are drawn about the relationship between their variables.

As with the previous analysis, there are no moderate or strong relationships in any of the guidelines. This means that no conclusions can be drawn about the relationship between variables or that said relationship is not decisive (see Fig. 12).

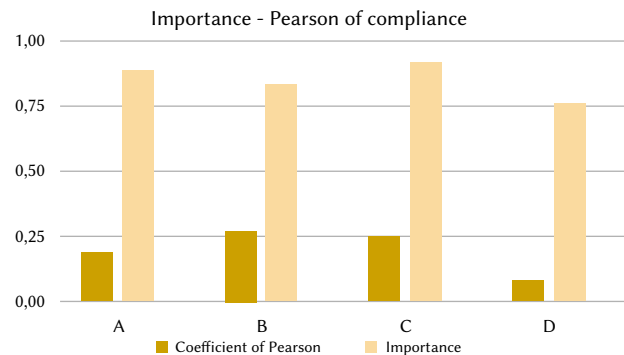


Fig. 12. Statistical analysis of the importance - Pearson of compliance.

In the study of these variables all possible relationships are positive, which would mean that if there was a relationship, it would be direct.

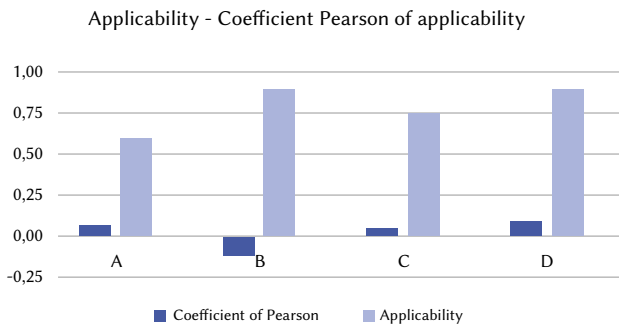


Fig. 13. Statistical applicability – coefficient Pearson of applicability.

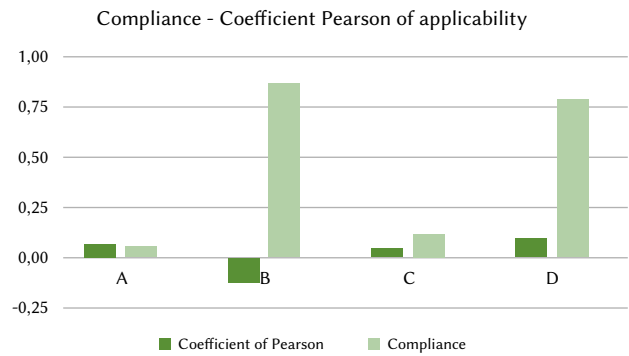


Fig. 15. Statistical compliance – coefficient Pearson of applicability.

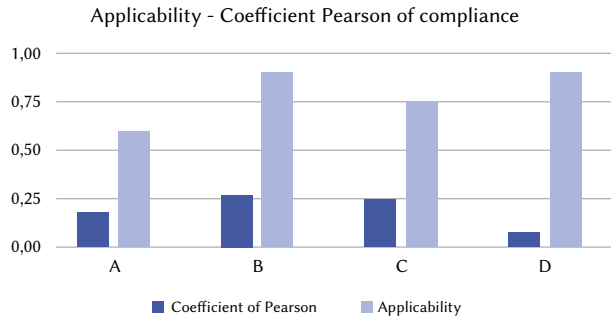


Fig. 14. Statistical applicability – coefficient Pearson of compliance.

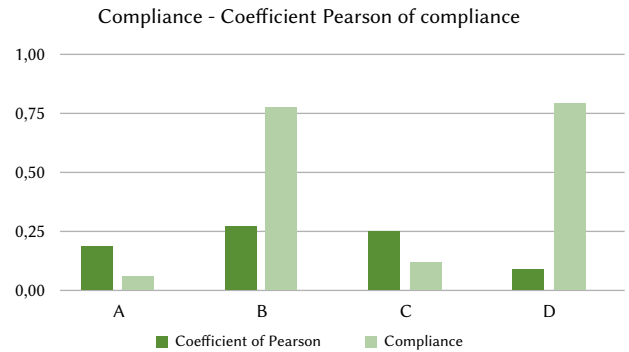


Fig. 16. Statistical compliance – coefficient Pearson of compliance.

That is, the greater the degree of importance attached to the guideline, there would be greater compliance. It should be remembered that compliance can only occur if the guideline has been applied. There are applied guidelines that have not been met, but not vice versa.

The most prominent relationship is that of guideline B: (Notice of response to actions). The previous analysis concluded that the more important, the less application. Now it is determined that the more important, the better compliance.

On the other hand, it was previously demonstrated that this is a very applied guideline, and it was also concluded that the importance was not very decisive in the comparison because there were generally high scores.

This could mean that although there is no relationship between variables, it is a well-applied and well-fulfilled guideline. In this example, importance is not decisive.

D. Analysis of the APPLICABILITY and COMPLIANCE

The relationship pattern between the *Pearson coefficient of applicability/compliance* and *Applicability* variables and *Pearson coefficient of applicability/compliance* and *Compliance* variables is analyzed. Conclusions are drawn about the relationship between their variables.

The basis of the Pearson coefficient is that the more intense the concordance (in the direct or inverse sense), the product gets more value. It measures the statistical relationship between two continuous variables. If the association between the elements is not linear, then the coefficient is not represented. This is the case of the relationships in this study. It has already been shown that there is no relationship between the variables (neither in applicability nor in compliance). Therefore, we cannot offer useful dispersion diagrams nor will the relationships be explained again in the following analyzes. However, we offer the graphs that compare the results of these coefficients with the Applicability and Compliance results of the previous point, Fig. 13-16.

It is necessary to emphasize that to obtain the Pearson coefficient results, the Importance variable was always analyzed, and it was compared alternately with the variable Application or Compliance.

This means that these figures represent on the one hand the relationship between variables, and on the other, the result of Applicability or Compliance. What is intended is that, although there is no relationship between the variables studied, useful conclusions can be drawn from this experiment.

VI. DISCUSSION

Fig. 17 presents the relationship between the results of the variables. The most applied recommendations are B: (Notice of response to actions) and D: (Descriptive information about errors). A: (Identify viewed or selected items) is the worst applied. The best complied with recommendations are also B: (Notice of response to actions) and D: (Descriptive information about errors). It could say that these two guidelines are innately applied because even their compliance is adequate. The worst complied with recommendations are A: (Identify viewed or selected items) and C: (“Undo” to go back on tasks sensitive). Curiously, C: (“Undo” to go back on tasks sensitive) is applied spontaneously by 75% of users. However, almost all comply with some errors.

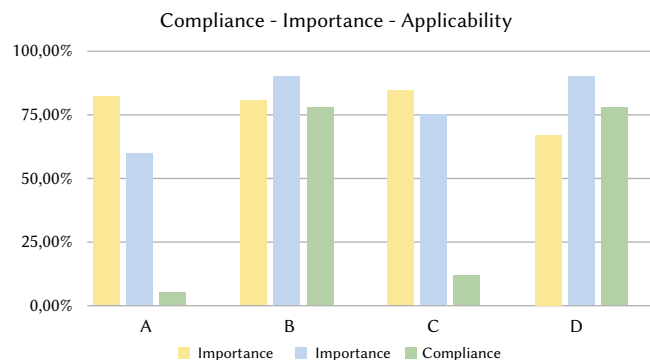


Fig. 17. Importance-compliance-applicability.

A. Best Rated Guidelines

The best-applied guideline is D: (Descriptive information about errors), with 90% application. The best-complied guideline is D: (Descriptive information about errors), with 78.5% compliance. It could be said that this recommendation is not a “necessary” reason for learning, because it seems to belong to the nature of web development.

The same could be said of recommendation B: (Notice of response to actions). It has 90% applicability and 77% compliance.

Although it is determined that the importance factor does not provide too much information for offering such high values, recommendation B: (Notice of response to actions) is considered very important. However, D: (Descriptive information about errors) is the least important recommendation, with 67%.

B. Worst Rated Guidelines

The worst applied guideline is A: (Identify viewed or selected items), with a 60% application. The worst complied guideline is A: (Identify viewed or selected items), with 6% compliance. This guideline needs, on the one hand, to be taught in usability agendas. Because it is the most unknown of the recommendations. Besides, on the other hand, it needs a lot of practice in training, because indeed all web developers have problems with errors. When web developers receive usability training, they give it 82% importance.

The next worst-performing address is C: (“Undo” to go back on tasks sensitive), with 12%. Unlike the A: (Identify viewed or selected items), the C: (“Undo” to go back on tasks sensitive) is applied by 75% of the participants. It could be said that this address, rather than being known, needs to be practiced by web developers.

VII. CONCLUSIONS

Our work aims to discover which usability guidelines related to “Responding to user actions” are considered more and less important for web developers, and also which are applied more and less. We tried to discover if there is a relationship between importance and application in these kind of web usability guidelines.

To meet this research objectives, we designed two experiments. First, a team of 20 web developers without knowledge of web usability designed 20 websites. All web developers were unaware of the web usability recommendations that would be analyzed later. The goal was to find out if the application of any of these guidelines is innate or should be learned. Second, web developers trained in web usability, specifically in guidelines related to “Responding to user actions”, and responded to two surveys: 1) It aimed to know the level of importance that web developers give to each of the guidelines, once studied. 2) It aimed to evaluate the websites to know the level of compliance with these guidelines.

Data analysed in this research work suggest that there is no relationship between importance and application in this kind of web usability guidelines. Web developers without usability training habitually made mistakes related to A: (Identity viewed or selected items). For instance, in the user interface, include an “Undo” option to allow users to revert sensitive actions or tasks. They made few mistakes related to B (Notice of response to actions). Instead, once they have training in usability, they consider very important C (“Undo” to return to sensitive tasks) just one of the guidelines that they least applied on their websites. The guideline as less important was D (Descriptive information about errors) which is the second most used guideline on their websites.

VIII. FUTURE RESEARCH LINES

We have published the research carried out about the Recommendations Group 1, 2, 3, 4 [8]-[10], [16] and this is the research of Group 5.

The next thing will be to test these results with more participants, including other questions of interest, for example, if there are patterns of behaviour while the web development and if this affects web usability, or if the application of the guidelines also depends on the area to which the website is intended.

We also seek to validate the improvement of each guideline in the user experience. This is interesting to support this saga of papers that we have already published. At this moment we are working on the development of a validation tool.

And we also want to compare the improvement offered by each guideline depending on the level of experience the user has.

REFERENCES

- [1] R. González Crespo, J. Pascual Espada, and D. Burgos, “Social4all: Definition of specific adaptations in Web applications to improve accessibility,” *Journal of Computers Standards and Interfaces*, vol. 48, pp. 1-9, 2016.
- [2] E. Bader, E. M. Schön, and J. Thomaschewski, “Heuristics Considering UX and Quality Criteria for Heuristics,” *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 4, no. 6, pp. 48-53, 2017.
- [3] J. Nielsen, “10 usability heuristics for user interface design,” *Nielsen Norman Group*, 1995.
- [4] A. Lodhi, “Usability Heuristics as an assessment parameter: For performing Usability Testing”; in *2010 2nd International Conference Software Technology Engineering*, vol. doi:10.1109/ICSTE.2010.5608809, pp. V2-256-V2-259, 2010.
- [5] F. Paz, C. Villanueva, C. Rusu, S. Roncagliolo, and J. Pow-Sang, “Experimental Evaluation of Usability Heuristics,” in *2013 10th International Conference Information Technology New Generations*, vol. doi:10.1109/ITNG.2013.23, pp. 119-126, 2013.
- [6] C. Mariage, J. Vanderdonck, and C. Pribeanu, “State of the Art” of *Web Usability Guidelines*, 2005, pp. 688-700.
- [7] L. Alonso-Virgós, et al., “Design specific user interface for people with Down syndrome using suitable WCAG 2.0 guidelines,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, no. 5, pp. 1359-1374, 2018.
- [8] L. Alonso-Virgós, J. Pascual Espada, and R. González Crespo, “Analysing compliance and application of usability guidelines on efficient and understandable controls,” *Computer Standards & Interfaces*, vol. 66, p. 103349, 2019.
- [9] L. Alonso-Virgós, J. Pascual Espada, and R. González Crespo, “Analyzing compliance and application of usability guidelines and recommendations by web developers,” *Computer Standards & Interfaces*, vol. 64, pp. 117-132, 2019.
- [10] L. Alonso-Virgós, L. Rodríguez Baena, J. Pascual Espada, and R. González Crespo, “Web Page Design Recommendations for People with Down Syndrome Based on Users’ Experiences,” *Sensors*, vol. 18, no. 11, p. 4047, 2018.
- [11] H. Purchase, J. Allder, and D. Carrington, “User preference of graph layout aesthetics: A UML study,” *International Symposium Graph Draw*, no. 5-18, 2000.
- [12] D. Green and J. Pearson, “Integrating website usability with the electronic commerce acceptance model,” *Behaviour & Information Technology*, vol. 30(2), pp. 181-199, 2011.
- [13] A. Hinderks, D. Winter, F. J. Domínguez Mayo, M. J. Escalona, and 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.
- [14] S. Majumder, S. Chowdhury, N. Dey, and K. C. Santosh, “Balance Your Work-Life: Personal Interactive Web-Interface,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 7, pp. 90-96, 2022.

- [15] M. Bernard, "Examining user expectations for the location of common e-commerce web objects," *Usability News*, vol. 4(1), pp. 1-7, 2002.
- [16] L. Alonso-Virgós and J. Thomaschewski, "Test usability guidelines and follow conventions. Useful recommendations from Web Developers," *Computer Standards & Interfaces*, p. 103423, 2020.
- [17] M. Schrepp, R. Otten, K. Blum, and J. Thomaschewski, "What Causes the Dependency between Perceived Aesthetics and Perceived Usability?," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, issue Regular Issue, no. 6, pp. 78-85, 2021.
- [18] J. Nielsen, "Design Guidelines for Homepage Usability," *Nielsen Norman Group*, 2001.
- [19] A. Baldominos, F. De Rada, and A. Sae, "DataCare: Big Data Analytics Solution for Intelligent Healthcare Management," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 4, no. 7, pp. 13-20, 2018.
- [20] J. Grobelny, W. Karwowski, and C. Drury, "Usability of Figure al icons in the design of human-computer interfaces," *International Journal of Human-Computer Interaction*, vol. 18(2), pp. 167-182, 2005.
- [21] M. Rauschenberger, S. Olschner, M. Cota, M. Schrepp, and J. Thomaschewski, "Measurement of user experience: A Spanish language version of the user experience questionnaire (UEQ)," *7th Iberian Conference on Information Systems and Technologies (CISTI 2012)*. *IEEE*, 2012.
- [22] J. Nielsen and H. Loranger, "Prioritizing Web Usability," *New Riders Publishing, Thousand Oaks, CA, USA*, 2006.
- [23] X. Wang, "Using Cognitive Walkthrough procedure to prototype and evaluate dynamic menu interfaces: A design improvement", in *2008 12th International Conference Computer Supported Cooperative Work Des.*, pp. 76-80, 2008.
- [24] J. Duan, "Research on visualization techniques for web usability analysis", in *2nd International Conference Information Science Engineering*, pp. 5366-5369, 2010.
- [25] P. Filip and L. Lukáš, "Webalyt: Implemetation of architecture for capturing web user behaviours with feedback propagation," *28th International Conference Radioelektronika (RADIOELEKTRONIKA)*. *IEEE*, 2018.
- [26] C. Li and C. Kit, "Web structure mining for usability analysis", in *2005 IEEE/WIC/ACM International Conference Web Intell*, pp. 309-312, 2005.
- [27] P. Kortum, "HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces", San Francisco, CA, USA: *Morgan Kaufmann Publishers Inc.*, 2008.
- [28] N. Mahyavanshi, M. Patil, and V. Kulkarni, "A realistic study of user behavior for refining web usability" in *2017 Int. Conf. I-SMAC (IoT Soc. Mobile, Anal. Cloud)*, pp. 450-453, 2017.
- [29] J. Schrepp and M. Hinderks, "Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S)," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, pp. 103-108, 2017.
- [30] K. Munim, I. Islam, M. Khatun, M. Karim, and M. Islam, "Towards developing a tool for UX evaluation using facial expression", *3rd International Conference on Electrical Information and Communication Technology*, pp. 1-6, 2017.
- [31] L. Martin, "A tool to estimate usability of Web 2.0 applications", in *11th IEEE International Symposium on Web Systems Evolution*, 2009," pp. 83-86, 2009.
- [32] MM, "Usability Testing Case Study: Objective Evaluation vs Subjective Evaluation," [Online]. Available: <https://medium.com/@nurisanendita/usability-testing-case-studies-objective-evaluation-vs-subjective-evaluation-b5e67d678e5e>.
- [33] C. Sik-Lányi, V. Szűcs, and T. Guzsvinecz, "Usability and colour-check of a healthcare WEB-site", in *2017 IEEE 30th Neumann Colloquium*, pp. 111-116, 2017.
- [34] J. Kirakowski and B. Cierlik, "Measuring the Usability of Web Sites," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 42, pp. 424-428, 1998.
- [35] T. Tullis, S. Fleischman, M. McNulty, C. Cianchette, and M. Bergel, "An Empirical Comparison of Lab and Remote Usability Testing of Web Sites," in *U.P.A. Conference (Ed.)*, *Usability Professionals Association*, 2002.
- [36] F. Paz, F. Paz, J. Pow-Sang, and L. Collantes, "Usability Heuristics for Transactional Web Sites," in *2014 11th International Conference Information Technology New Generations*, doi:10.1109/ITNG.2014.81, 2014.
- [37] D. Quiñones, C. Rusu, and S. Roncagliolo, "Redefining usability heuristics for transactional web applications," *2014 11th International Conference on Information Technology: New Generations. IEEE*, 2014.
- [38] S. Papaloukas, K. Patriarcheas, and M. Xenos, "Usability Assessment Heuristics in New Genre Videogames" in *2009 13th Panhellenic Conference on Informatics*, doi:10.1109/PCI.2009.14, pp. 202-206, 2009.
- [39] L. Gregory; S. Brent; Z. Nida, "Use of product viewing histories of users to identify related products." *U.S. Patent* No 6,912,505, 28 Jun. 2005.
- [40] B. Dominic. "Predicting user response to advertisements." *U.S. Patent Application* No 12/410,400, 8 Abr. 2010.
- [41] G. Salvador Cobos, M.D. Cima Cabal, F. Machío Regidor and L. Alonso Virgós. "Cyber-Physical System Architecture for Minimizing the Possibility of Producing Bad Products in a Manufacturing System" in *Innovative Design and Operation of Digital Manufacturing Equipment-Trends and Prospect of Manufacturing Intelligence*. Intech Open, 2019.
- [42] F. Eelke; B. Jan. "Architecturally sensitive usability patterns." *Department of Mathematics and Computing Science, University of Gronigen, Netherlands*, 2003, pp. 1-19.
- [43] H. Zhao; B. Morad, "Usability and credibility of e-government websites." *Government Information Quarterly*, 2014, vol. 31, no. 4, pp. 584-595.



Lucía Alonso Virgós

Lucía Alonso Virgós is a computer scientist with interest in web usability. Professor and Director of Industrial and Electronic Area of de International University of La Rioja.



Jordán Pascual Espada

Jordán Pascual Espada is a computer scientist with an interest in web usability; he is a Deputy Director of the Engineering School of University of Oviedo.



Gustavo Rossi

Ph.D. from PUC-RIO Brazil University. Full professor at the Computer Science College at Plata National University and head of the LIFIA research group. Visiting professor at the Universities of Lyon and Montpellier in France, received the Habilitation pour Diriger Recherches (HDR) at INSA-Lyon. Part of the PC committee of many significant conferences like ACM WWW, ICWE, and ACM Hypertext. He has published more than 200 research papers, most of them in the field of modeling and design issues for advanced Web Applications, particularly those that involve context-aware behaviors.

