

“Some people worry that artificial intelligence will make us feel inferior, but then, anybody in his right mind should have an inferiority complex every time he looks at a flower.”

Alan Kay

EDITORIAL TEAM

Editor-in-Chief

Dr. Elena Verdú, 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

Dr. Yamila García-Martínez Eyre, Universidad Internacional de La Rioja (UNIR), Spain

Office of Publications

Editorial Coordination

Lic. Ángela Porras, Universidad Internacional de La Rioja (UNIR), Spain

Indexing and Metrics

Dr. Álvaro Cabezas, 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

Advisory Editors

Dr. David Camacho, Technical University of Madrid, Spain

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

Dr. Robertas Damaševičius, Kaunas University of Technology, Lithuania

Dr. Gwanggil Jeon, Incheon National University, South Korea

Dr. Yiu-Ming Cheung, Hong Kong Baptist University, Hong Kong

Associate Editors

Dr. Kuan-Ching Li, Providence University, Taiwan

Dr. Miroslav Hudec, VSB - Technical University of Ostrava, Czech Republic

Dr. Mahdi Khosravi, 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. Qin Xin, University of the Faroe Islands, Faroe Islands, Denmark

Dr. Yaping Mao, Qinghai Normal University, China

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. Amrit Mukherjee, University of South Bohemia, Czech Republic

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

Dr. Wensheng Gan, Jinan University, China

Dr. Francesco Piccialli, University of Naples Federico II, Italy

Dr. Chang Choi, Gachon University, South Korea

Dr. Junxin Chen, Dalian University of Technology, China

Dr. Richard Chbeir, Université de Pau et des Pays de l'Adour, France

Dr. Hao-Tian Wu, Guangzhou University, China

Dr. Patrick C. Hung, Ontario Tech University, Canada

Dr. Mahendra Deore, MKSSS's Cummins College of Engineering for Women, India

Dr. Ting Cai, Hubei University of Technology, China

Dr. Andre de Lima Salgado, Universidade Federal de Lavras, UFLA, Brazil

Dr. Hsiao-Ting Tseng, National Central University, Taiwan

Dr. Mengke Li, Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), China

Editorial Board Members

Dr. Rory McGreal, Athabasca University, Canada

Dr. Óscar Sanjuán Martínez, Lumen Technologies, USA

Dr. Nilanjan Dey, Techno International New Town, India

Dr. Anis Yazidi, Oslo Metropolitan University, Norway

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

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. Omer Melih Gul, Istanbul Technical University (ITU), Turkey

Dr. S. Vimal, KIT-Kalaingar Karunanidhi Institute of Technology, Coimbatore, India

Dr. Jordán Pascual Espada, Oviedo University, Spain

Dr. Ioannis Konstantinos Argyros, Cameron University, USA

Dr. Palanichamy Naveen, Dr. N.G.P. Institute of 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. B. Cristina Pelayo G. Bustelo, University of Oviedo, 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. Masao Mori, Tokyo Institute of Technology, Japan

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. David L. La Red Martínez, National University of North East, Argentina

Dr. José Estrada Jiménez, Escuela Politécnica Nacional, Ecuador

Dr. Moamin A. Mahmoud, Universiti Tenaga Nasional, Malaysia

Dr. Madalena Ribeiro, Polytechnic Institute of Castelo Branco, Portugal

Dr. Edward Rolando Núñez Valdez, University of Oviedo, Spain

Dr. Juha Rönning, 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. Pranav Gangwani, Florida International University, Miami, USA

Dr. Fernando López, Universidad Complutense de Madrid, Spain

Dr. Runmin Cong, Beijing Jiaotong University, China

Dr. Abel Gomes, University of Beira Interior, Portugal

Dr. Mohammad Javad Ebadi, Chabahar Maritime University, Iran

Dr. Andreas Hinderks, University of Sevilla, Spain

Editor's Note

Are We on the Way to Become Machines From Our Own Machines?

ARTIFICIAL Intelligence (AI) is a scientific discipline that aims to drive disruptive scenarios for science-based technical developments that solve complex problems. The IJIMAI journal's scope is precisely to demonstrate how the combination of two factors – technical foundations and sought-after applications – must guide future AI developments to find solutions to complex real-world problems.

This IJIMAI publication opens with an article that considers the current framework for AI fundamentals: **how can we improve AI technology to find solutions to real-unsolved problems?** The initial answer seems to be related with a desired self-consistent procedure: let machines learn from our experience. In the article by Alotaibi et al., the analysis of neural networks in terms of the parameters used, how they work, and how do they respond to the problem itself led the authors to a rationale for decision-making regarding the performance of different neural models.

The immediate question that arises is whether there are any universal and fundamental criteria that can be used to define the models that guide AI methods. Apparently, there are not such universal methods, and we are faced with a challenging open problem. Subsequent manuscripts will provide readers with more in-depth insights into this issue.

The next articles place you at the forefront of current AI challenges, while using AI as a medium to assess objective results (understood as numerical evidence) to shed light on current problems arising from clearly defined, real-world areas of interest. The common factor is the complex nature of the issues at hand. This complexity is derived from the large amount of data and the development of a rationale to comprehend the interactions between them, i.e. data mining research (DM).

Real-world problems vary from highly demanding social science scenarios (where the huge number of variables involved is undeniable) to the most purely nature-inspired scenarios (where it is assumed that some well-defined laws should govern, reducing the freedom associated with the large number of natural variables). Both scenarios would probably find a route to further success by incorporating social rules and/or natural laws into the methodologies used to develop their associated AI algorithms. Such inclusions are essential for supporting the rationale behind what is known as machine learning (ML), a point also highlighted in this publication. From this point onwards, and in this IJIMAI edition, a variety of technical areas, including image processing, climate modelling, language processing and emotion analysis, are used to test current AI models.

The second contribution of this IJIMAI number is dedicated to numerical methods applied to image processing. Extracting information from image processing is one of the most challenging problems in information technology scenarios. Díaz-Pacheco et al. present methodological results that encourage the use of deep learning approaches for supervised (discriminative) learning and guide further unsupervised (generative) learning. The use of such results to inform decisions in the tourism sector is presented as a real-world application.

The next contribution deals with climate evolution predictions to create more robust forecasting systems. Such challenges rely on the need for more general and demanding fluid mechanics laws and complex systems analysis. Despite the progress made in these areas,

the effective numerical integration of climate model projections with measured rainfall data remains challenging. The relevance of such analysis for vital sectors such as disaster prevention and agriculture-based industries is beyond doubt. Considering the results presented by Oswalt et al. would therefore be a meaningful way to combine advanced AI models with data fusion techniques, thereby enhancing the accuracy and reliability of rainfall predictions.

The articles by Paramasivam et al. and Zhang et al. are closely related. The manuscript presented by Paramasivam et al. deals with speech emotion recognition and the impact of data acquisition on the subsequent design of processing algorithms for machine learning-based solutions. The work presented by Zhang et al. deals with sentiment analysis in relation to AI use, which is probably a technological step further than the previously presented AI for 'language analysis'. The relevance of the topics addressed in both studies is widely recognised and encouraged by researchers who specialise in leveraging the strengths of AI to advance areas related to social behaviours. Examples include personalised healthcare, affective computing and human-machine interactions, as well as the omnipresent market insertion and interferences in our daily routines, which have been conditioned by commercial interests.

While the huge diversity of fields claiming methodological applications of AI is clear from the articles presented so far, the next three contributions focus on more specific technical concerns. Due to their relevance, such technical issues must be included as part of well-defined branches within AI research: the aforementioned 'data mining' and 'machine learning'.

The manuscript presented by Cartensen et al. could be considered a piece of research highlighting the importance of DM. The analysis presented therein once again raises the question regarding the need for fundamental rules to make data analysis more valuable in terms of setting guarantees for further algorithm designs. Successful contrast with selected experimental training data does not offer strong enough guarantees for AI design. However, presenting this AI problem-solving route is a valuable result, as it puts us at the forefront of state-of-the-art AI technologies.

The next two articles could be highlighted as contributions that are more closely related to the use of AI for ML designs. The first, by Suruliandi et al., uses what could be considered ML algorithms to perform tasks related to the well-known problem of healthcare assistance. In this case, the authors use clinical data collected for the evaluation of possible cases of coronavirus as training sets. The article then goes on to discuss the optimisation of effective diagnostic algorithms. The other manuscript, by Guo et al., proposes a mechanism that improves the state of the art in image analysis scenarios by optimising network architectures. The work presented goes into person-re-identification control mechanisms.

The last four manuscripts in this issue of IJIMAI review the issues that guide the scope of this journal and of AI research in general. They highlight the most challenging unsolved problems and the need to conscientiously improve AI technologies.

Martínez Núñez et al. analyse the impact of deep learning methods combined with visual computing techniques to address constraints and enhance operational capabilities around railway tracks. Tejero et al. address the universal problem of treating complex systems derived

from the analysis of social events. However, the most notable concept studied and evaluated here is misinformation. There is no doubt that this concept is closely related to the integration of AI methods into societal events. The relevance of this discussion in the current context must therefore be highlighted. **Is misinformation inherent to AI at this stage?** The question remains open.

The potential applications of AI methods in predicting, modulating and controlling human behaviours are far beyond current estimates. There is no doubt that AI technologies are already exerting control over interactions involving at least one human element. This is a hopeful and risky promise. With these factors in mind, trial cases are studied in next articles, one affecting the economic sector and the other concerning educational scenarios.

The economy sector is inevitably modulated by client and market interactions. The question of how these interactions can be controlled to generate profit from AI results is analysed by Kollmorgen et al. The second trial case involves students and their educational context. Based on similar technological foundations, the integration of AI into educational contexts is opening up new opportunities and presenting new challenges. Sánchez-Canella et al. intend to provide guidelines to improve interactions between AI and the educational final targets: students. **Is AI in education recommended? Is it useful? Is it challenging?** Time is needed.

Once again, this IJIMAI edition supports efforts to guide researchers in a collaborative endeavour to establish robust methodological frameworks that underpin the conscious application of AI techniques to real-world problems. The lack of systematic methodologies for designing self-guided algorithms -ML- and for defining proper learning databases -DM-, highlights the need to consciously prioritise basic research in order to formalise and control the rules governing the behaviour of complex systems under analysis.

Finally, I would like to thank all the authors who contributed to this edition of IJIMAI for their valuable contributions. I encourage them and all the other researchers involved to use their expertise to ensure the safe and successful advancement of AI.

To conclude this letter, I would like to invite you to consider the following questions: **should AI learn from our presumed natural intelligence? If so, is it a good idea?**

Dr. Yamila García-Martínez Eyre i Canals
Managing Editor
Universidad Internacional de La Rioja

TABLE OF CONTENTS

EDITOR'S NOTE.....	3
PERFORMANCE AND COMMUNICATION COST OF DEEP NEURAL NETWORKS IN FEDERATED LEARNING ENVIRONMENTS: AN EMPIRICAL STUDY	6
MEASURING THE DIFFERENCE BETWEEN PICTURES FROM CONTROLLED AND UNCONTROLLED SOURCES TO PROMOTE A DESTINATION. A DEEP LEARNING APPROACH	18
AN ADAPTIVE SALP-STOCHASTIC-GRADIENT-DESCENT-BASED CONVOLUTIONAL LSTM WITH MAPREDUCE FRAMEWORK FOR THE PREDICTION OF RAINFALL.....	32
A ROBUST FRAMEWORK FOR SPEECH EMOTION RECOGNITION USING ATTENTION BASED CONVOLUTIONAL PEEPHOLE LSTM.....	45
OPTIMAL TARGET-ORIENTED KNOWLEDGE TRANSPORTATION FOR ASPECT-BASED MULTIMODAL SENTIMENT ANALYSIS	59
TKU-PSO: AN EFFICIENT PARTICLE SWARM OPTIMIZATION MODEL FOR TOP-K HIGH-UTILITY ITEMSET MINING	70
PREDICTION OF COVID-19 USING A CLINICAL DATASET WITH MACHINE LEARNING APPROACHES	82
MULTISCALE ATTENTIONAL SQUEEZE-AND-EXCITATION NETWORK FOR PERSON RE-IDENTIFICATION.....	99
AUTOMATIC SURVEILLANCE OF PEOPLE AND OBJECTS ON RAILWAY TRACKS.....	107
COMBATING MISINFORMATION AND POLARIZATION IN THE CORPORATE SPHERE: INTEGRATING SOCIAL, TECHNOLOGICAL AND AI STRATEGIES.....	117
SELECTING THE APPROPRIATE USER EXPERIENCE QUESTIONNAIRE AND GUIDANCE FOR INTERPRETATION: THE UEQ FAMILY	126
PLATFORM FOR IMPROVING THE USER EXPERIENCE IN THE CREATION OF EDUCATIONAL MULTIPLAYER VIDEO GAMES	140

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 © 2025 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/>

Performance and Communication Cost of Deep Neural Networks in Federated Learning Environments: An Empirical Study

Basmah K. Alotaibi^{1,2}, Fakhri Alam Khan^{1,3,4*}, Yousef Qawqzeh⁵, Gwanggil Jeon⁶, David Camacho⁷

¹ Information and Computer Science Department, King Fahd University of Petroleum and Minerals, Dhahran 31261 (Saudi Arabia)

² Department of Computer Science, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 13318 (Saudi Arabia)

³ Interdisciplinary Research Centre for Intelligent Secure Systems, King Fahd University of Petroleum and Minerals, Dhahran (Saudi Arabia)

⁴ SDAIA-KFUPM Joint Research Center for Artificial Intelligence, King Fahd University of Petroleum and Minerals, Dhahran (Saudi Arabia)

⁵ College of Information Technology, Fujairah University (UAE)

⁶ Department of Embedded Systems Engineering, Incheon National University, 119 Academy-ro, Yeonsu-gu, Incheon, 22012 (Korea)

⁷ Department of Computer Systems Engineering Universidad Politécnica de Madrid Madrid (Spain)

* Corresponding author: fakhri.khan@kfupm.edu.sa

Received 14 May 2024 | Accepted 6 September 2024 | Early Access 16 December 2024



ABSTRACT

Federated learning, a distributive cooperative learning approach, allows clients to train the model locally using their data and share the trained model with a central server. When developing a federated learning environment, a deep/machine learning model needs to be chosen. The choice of the learning model can impact the model performance and the communication cost since federated learning requires the model exchange between clients and a central server in several rounds. In this work, we provide an empirical study to investigate the impact of using three different neural networks (CNN, VGG, and ResNet) models in image classification tasks using two different datasets (Cifar-10 and Cifar-100) in a federated learning environment. We investigate the impact of using these models on the global model performance and communication cost under different data distribution that are IID data and non-IID data distribution. The obtained results indicate that using CNN and ResNet models provide a faster convergence than VGG model. Additionally, these models require less communication costs. In contrast, the VGG model necessitates the sharing of numerous bits over several rounds to achieve higher accuracy under the IID data settings. However, its accuracy level is lower under non-IID data distributions than the other models. Furthermore, using a light model like CNN provides comparable results to the deeper neural network models with less communication cost, even though it may require more communication rounds to achieve the target accuracy in both datasets. CNN model requires fewer bits to be shared during communication than other models.

KEYWORDS

Communication Cost, Convolutional Neural Network (CNN), Deep Neural Networks, Distributive Learning, Federated Learning, Neural Network, Performance, Residual Neural Network (ResNet), Visual Geometry Group (VGG).

DOI: 10.9781/ijimai.2024.12.001

I. INTRODUCTION

THE expansion of information and communication technology has increased the availability of data and computing resources, resulting in the Big Data era. This increasing data generated in the network requires efficient knowledge extraction and processing mechanisms to benefit from it. The data generated can be utilized as training data

to provide the edge devices in the network with intelligence. However, traditional machine/deep approaches necessitate the collection of data to a central location to train the model and extract knowledge from it. Collecting the data to a central location can cause a significant transmission delay and raise privacy concerns due to sharing some private information through the network. Therefore, traditional machine learning approaches that require data collection in a central

Please cite this article as: B. K. Alotaibi, F. A. Khan, Y. Qawqzeh, G. Jeon, D. Camacho. Performance and Communication Cost of Deep Neural Networks in Federated Learning Environments: An Empirical Study, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 4, pp. 6-17, 2025, <http://dx.doi.org/10.9781/ijimai.2024.12.001>

location face various challenges, such as network communication and data privacy [1]. To tackle these issues, Federated Learning (FL) was introduced [2] to allow the model to be trained locally at the network edge and share the learned model rather than the data. FL is a distributive collaborative learning process that allows devices (known as clients) to train the model locally using their local data and share the trained model with a central server. The central server aggregates the received local models to create the global model. The learning process in federated learning is performed in rounds where, in each round, the central server provides the global model to the clients to train the model locally using their data [2].

Federated learning technology offers numerous advantages over traditional learning methods. It provides for more effective usage of network bandwidth and protects data privacy, as raw data is not demanded to be transferred to the server. Moreover, federated learning can employ the computing resources and diverse datasets on clients' devices to enhance the quality of the global model [3], [4], [5]. With these benefits, federated learning can be applied in various areas such as healthcare, transportation, IoTs, and mobile applications (such as next-word prediction) [6], [7].

However, federated learning faces various challenges because of its decentralized approach such as quality of the training data, the distributed architecture, the type of devices used to train the model, and the communication and aggregation mechanisms used; which affect the learning process in FL [8]. The clients' data in FL is non-Independent and Identically Distributed (non-IID) data due to varying device usage and location, as each client's data is dependent on their device usage and location. This means that the assumption of the IID data used in machine learning algorithms cannot be applied in FL. Therefore, FL encounters the additional challenge of data heterogeneity [4], [9]. Furthermore, during FL training, the clients and the central server exchange the local and global models in multiple rounds. However, this communication process can be a bottleneck due to the network's limited resources [10], [11]. The devices in FL vary in their computational power, storage, and network connectivity, and this heterogeneity could unbalance the training time and affect global model training [12], [13], [14].

Despite federated learning's potential, most research has focused on overcoming challenges like communication efficiency, data heterogeneity, and privacy preservation. However, an overlooked aspect of FL is the influence of various deep learning models on the overall performance and efficiency of FL. Understanding the impact of different neural network architectures within FL is crucial for optimizing model performance and communication resource efficiency.

In contrast to the traditional centralized learning approach, FL perform the training in distributed manner on the clients' devices and shares the local models with the central server through the network for several rounds. However, sharing these local models can be costly when using a deep neural network with many parameters. For that, this study aims to investigate the impact of using different neural networks on the model performance and communication cost, focusing on image classification tasks. Specifically, the research intends to evaluate and analyze the performance of a Convolutional Neural Network (CNN) model along with two complex variations of it, namely Visual Geometry Group (VGG) and Residual Neural Network (ResNet), which are widely used by the researchers when evaluating their proposed work in image classification tasks [15], [16], [17], [18], [19]. These three models are mainly used with Cifar-10 and Cifar-100 datasets [20]. In this study, we aim to address the following research question: Do we need a deeper network in a federated learning environment? by studying the performance of VGG-11, ResNet-18, and comparing them with a lighter CNN model, that is the same model

used in the study that proposed the FL approach [2]. We aim to gain a deeper understanding of the benefits and drawbacks of using these models in FL environment.

The main contributions of this study are as follows:

1. Conducting an empirical study that investigates the impact of utilizing three different neural networks (CNN, ResNet-18, and VGG) for image classification task in a FL environment using two widely recognized datasets (Cifar-10 and Cifar-100).
2. We performed a comparative experiment and analyzed the performance of these three models under different data distribution settings (IID and non-IID), providing valuable insights into their behavior in different FL settings.
3. We studied and analyzed these models' performance and associated communication costs with different batch sizes and epoch values.
4. We provide insights into the trade-offs between model accuracy and communication efficiency.
5. Our findings shed light on the suitability of each neural network for FL, enabling researchers and practitioners to make informed decisions when selecting a learning model for their work.
6. To best of our knowledge, this study is the first to explore how the choice of different neural network models impacts the performance of federated learning.

The remainder of this work is organized as follows: Section II presents the literature review and highlight the datasets and local models in image classification tasks in federated learning. Section III show the system model. In section IV, we show the neural networks design, and the experiment settings. Section V shows the experiments results, while section VI shows the discussion. The conclusion and future work are provided in section VII.

II. LITERATURE REVIEW

Numerous studies have attempted to tackle the FL challenges using various techniques. For instance, Zhong *et al.* [21] uses a hierarchical clustering algorithm to overcome the non-IID data challenge by clustering the clients based on their model similarity and merge similar clusters. While Wu *et al.* [22] addresses communication cost and non-IID data challenges by using a threshold value to determine the importance of the local model to be uploaded to the server or skipped. Other studies [23], [24], [25] focus on non-IID data challenges by aiming to reduce client drift using different techniques such as decoupling and correction the local drift, rescaling the gradients, primal-dual variable that can adapt to data heterogeneity. The studies in [15], [16], [17] address communication challenges by reducing the number of bits exchanges using different compression techniques, such as Quantization and Count Sketch. The work in [18] improves the communication efficiency of FL by parallelizing the communication with computation to cover the communication phase with the training phase. Asynchronous technique is to overcome the communication bottleneck in FL [26], and another work uses partial synchronous technique to accelerate the training process in FL over the two-tier network using relay nodes to aggregate the model partially and reduce the communication rounds required [19]. The work of Li *et al.* aims to tackle the diversity in the computational capacity between devices and avoid waiting for slow devices by approximating the optimal gradients with a complete local training model using the Hessian estimation method to achieve the approximation, based on the heterogeneous local updates that has been received [27]. Another work uses a tier approach to overcome the latency caused by slow clients, by grouping clients into the same tier based on their response time to overcome the system heterogeneity [28]. Zeng *et al.* [29] proposes an energy-efficient bandwidth allocation and client selection scheme. The work

aims to reduce energy consumption while maximizing the selection of clients participating by adapting to both channel states and device computation capability when selecting clients and allocating the channel for them. Jebreel *et al.* [30] propose a mechanism to overcome the label-flapping attack in the federated learning environment to overcome malicious clients flipping their labels to poison the global model. Their work clusters clients based on their gradient parameters and analyzes the clusters to filter any potential threat. A novel backdoor attack in FL is introduced by Zhang *et al.* [31]. Their approach enables the attacker to optimize the backdoor trigger using adversarial training to enhance its persistence within the global training dynamics. In their work, they study the performance of existing techniques to overcome their attack and show their limitations.

These studies addressing FL challenges commonly use image classification tasks as their application when evaluating their proposed methods with different learning models. According to [20], [32], the most widely used datasets to test the model performance in FL are image datasets, and image classification tasks being the most commonly employed applications in FL. Furthermore, the study in [4] indicates that image datasets are the most used in FL.

CNN learning model is used in the study that propose FL [2] to evaluate its performance with MNIST [33] dataset. Many studies [21], [23], [26] use the same learning model when evaluating their proposed work, on different datasets such as CIFAR-10, CIFAR-100, and MNIST. Others have opted for a deeper neural network, such as VGG and ResNet, mainly when using Cifar-10 and Cifar-100 as their training dataset [15], [16], [17].

Table I shows various learning models and the datasets used to evaluate the performance in the FL environment. The table highlights that deeper network such as VGG and ResNet used different datasets such as Cifar-10 and Cifar-100, which include colored images, unlike

the MNIST datasets that are gray-colored images and widely used with simpler learning models. However, some research also uses Cifar-10 and Cifar-100 with a simpler model, such as CNN. The table indicates that CNN, ResNet, and VGG are commonly used with Cifar-10 and Cifar-100 datasets. Therefore, this study aims to investigate the performance of the three aforementioned learning models on the Cifar-10 and Cifar-100 datasets in a federated learning environment.

TABLE I. LEARNING MODELS AND DATASETS USED AT DIFFERENT STUDIES IN FEDERATED LEARNING ENVIRONMENT

Model	Dataset	Ref.
CNN	Cifar10	[21], [23], [25], [18], [26], [28]
	Cifar-100	[23]
	MNIST	[21], [23], [25], [28], [29], [30]
	Fashion MNIST	[25], [26], [28]
ResNet	Cifar-10	[22], [15], [16], [17], [18], [19], [27], [30], [31]
	Cifar-100	[15], [17], [18], [26], [27]
	FEMNIST	[17], [31]
	Tiny-ImageNet	[23], [31]
VGG	Cifar-10	[22], [15], [18], [24]
	Cifar-100	[22], [24]
MLP	MNIST	[16], [18]
	Fashion MNIST	[27]
Logistic regression	MNIST	[19]
	Cifar-10	[19]

Despite the extensive research addressing various challenges in federated learning, the impact of different learning models on federated learning performance has not been thoroughly investigated. Table I shows that some studies utilized more than learning model, however these models were utilized with different datasets. As shown in the Table II, the studies did not compare the impact of different

TABLE II. SUMMARY OF THE LITERATURE STUDIES

Ref.	Focus	Methodology	No. of Models for Same Dataset	Comparison between models	Hyper-parameter Tuning (Epoch-Batch)
[21]	Non-IID	Hierarchical clustering algorithm	1	X	X
[22]	Communication cost and Non-IID	Select model update	2	X	X
[23]	Non-IID	Decoupling and correcting local drift	1	X	X
[24]	Non-IID	Rescaling the gradient	1	X	X
[25]	Non-IID	Primal-dual variable to adapt to data heterogeneity	1	X	X
[15]	Communication cost	Compression	2	X	X
[16]	Communication cost	Compression	1	X	X
[17]	Communication cost	Compression	1	X	X
[18]	Communication efficiency	Parallelizing communication with computation		X	X
[26]	Communication bottleneck	Asynchronous technique	1	X	X
[19]	Accelerating training process	Partial synchronous technique using relay nodes to aggregate the model partially		X	X
[27]	Computational capacity	Approximating the optimal gradients with a complete local training model	1	X	X
[28]	Latency	Tier approach	1	X	X
[29]	Energy	Select client based on device computation capability and channel states	1	X	X
[30]	Security	Cluster clients based on gradients parameter and filter any potential threat	1	X	X
[31]	Security	Optimize attack trigger through an adversarial adaptation loss	1	X	X
Ours	Impact of Learning Models in Federated Learning	Evaluation of various deep learning models	3	✓	✓

deep learning models on the performance of federated learning. Even when multiple models are utilized, they are often used with different datasets, which makes direct comparisons difficult. Additionally, these studies did not thoroughly investigate the effects of hyper-parameter tuning, such as varying epochs and batch sizes, on model performance and communication efficiency. Selecting the learning model is essential to federated learning, as it impacts both model performance and communication cost. Therefore, an evaluation that compares multiple neural networks on the same datasets while considering hyper-parameter variations is important and is the focus of this study. Our study fills these gaps by evaluating the performance of various deep learning models within a Federated Learning framework, providing a unique contribution to the existing body of knowledge.

III. SYSTEM MODEL

In this section, we will provide a detailed explanation of the system model used in our study. This includes the principles and mechanisms of FL, the utilized aggregation algorithm, and the network architecture that we used.

A. Federated Learning

FL is a distributed collaborative learning process that was proposed by Google researchers in 2016 [2]. It is different from distributed (on-site) learning in that, in the latter, the central server provides the clients with an initial or pre-trained model, which the clients use to train their personalized models using their data. In this type of learning approach, there is no sharing of data or information [8], [34], [35].

In FL, there is a fixed set of Clients C , where each client c has its own datasets d_c , and at each round a fraction R of the clients C is selected to participate in this round to train the model [2]. Fig. 1 illustrates the FL architecture, where the central server sends the initial model to the participating clients. These clients then use this model to train a local model using their dataset. Afterward, the clients send the trained models to the server. The server then aggregates all the received local models using an aggregation mechanism. The process will be repeated for several rounds until a target is reached [2]. Typically, FL aims to minimize the objective function shown in (1):

$$\min_{\omega} F(\omega), \text{ where } F(\omega) := \sum_{c=1}^C p_c F_c(\omega) \quad (1)$$

Where C is the total number of clients, $p_c \geq 0$ and $\sum_c p_c = 1$, the p_c term define the impact of each client on the global model, where there are two natural settings existing which are: $p_c = \frac{1}{d}$, or $p_c = \frac{d_c}{d}$, where d represents the total data sample of all clients and d_c represents the data sample for client c , and F_c is the local objective function of client c [2], [7].

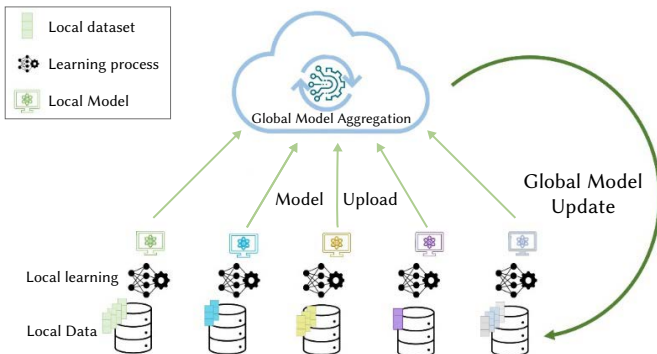


Fig. 1. The federated learning architecture.

In FL, the central server plays a vital role in providing an initial model, receiving updated local models from participating clients, aggregating these received local models, and subsequently disseminating a new global model to the participating clients. The most commonly used aggregation scheme in this type of learning is called Federated Averaging (FedAvg). FedAvg involves averaging the local stochastic gradient descent (SGD) updates. This method is usually implemented in a few general steps as follows [2], [7], [8]:

1. The server sets up the initial global model.
2. The server selects the participating clients (R, C), and sends the global model to them.
3. The clients that receive the global model train the received model using their local dataset. The most used technique is using SGD to compute the update.
4. The clients train the model for some epochs and upload the trained local model to the server.
5. The server then aggregates all the received local models using an averaging aggregation mechanism based on the clients' dataset size to create a new global model.
6. The steps from 2 to 5 are repeated for several rounds until a predefined target is reached.

Algorithm 1 (FedAvg):

The C Clients are indexed by c .

η : learning rate.

B : Batch size.

E : Local epochs.

Server executes:

Initialize the initial model ω_0

for each round $t = 1, 2, 3, \dots$ do

$m \leftarrow \max(R.C, 1)$

$C_t \leftarrow (\text{Select random set of } m \text{ clients})$

for every client $c \in C_t$ in parallel do

$\omega_{t+1}^c \leftarrow \text{ClientUpdate}(c, \omega_t) \dots$

$m_t \leftarrow \sum_{c \in C_t} n_c$

$\omega_{t+1} \leftarrow \sum_{c \in C_t} \frac{n_c}{m_t} \omega_{t+1}^c$

ClientUpdate(c, ω): // run on client c

$\beta \leftarrow \text{Spilt client dataset into batches of size } B$

for each local epoch i from 1 to E do

for batch $b \in \beta$ do

$\omega \leftarrow \omega - \eta \nabla l(\omega; b)$

return ω to the server

Algorithm 1 illustrates the process of the FedAvg algorithm where the (Server executes) section shows the steps that the server performs. In contrast, the (CleintUpdate) section illustrates the process that are performed at each client. The clients train a local model using a deep/machine learning approach and send the locally trained model to the server.

B. Network Architecture

The network design used in this study for FL is a centralized architecture. In this architecture, there is a central server S and a set of clients C , where each client c has its own local dataset d_c . The central server is responsible for initializing the global model and selecting the participating clients ($R.C$) from the clients set C . For this work, the central server selects the participating clients randomly.

In this network, each participating client c_i receives the global model from the central server and trains it locally using its own dataset d_c for a predefined local epoch. After the completion of training, the local model is shared with the central server for global aggregation. The central server applies the FedAvg aggregation scheme to aggregate the local models received from all participating clients. Then, the process of selecting clients and providing the global model is repeated. The process will continue for several communication rounds. The FL architecture used in this work is shown in Fig. 1.

IV. METHODS

This section presents the local learning model implemented by the federated learning clients and the experimental settings. It covers the local model architecture and the specific experimental parameters employed.

A. Local Model Design

In this subsection, we will discuss three deep learning models, namely CNN, ResNet, and VGG, that are utilized in a federated learning environment for image classification tasks. We will provide a general overview of their concepts and architecture in this subsection.

1. Convolutional Neural Network

CNN is a deep learning approach that can be utilized for tasks such as speech recognition and computer vision [36], [37]. CNN typically comprises three primary types of layers: convolutional, pooling, and fully connected layers. The convolutional layer is used for extracting features from the data. The pooling layer, on the other hand, reduces the size of the output from the convolutional layer and combines similar features to avoid redundancy. Finally, the fully connected layer establishes connections between the previous layer's output and the subsequent layer's input [38], [39], [40]. There are different well-known CNN architectures such as LeNet, AlexNet, GoogleNet, ResNet, VGG, which differ in terms of the used layers, the number of layers, the activation function used, and other factors [41], [38], [37]. The CNN model architecture used in this study is adopted from [2] and includes two convolutional layers, each followed by a max pooling layer, a fully connected layer, with ReLU activation function, and final SoftMax output layer.

2. Visual Geometry Group

The VGG model is a deep convolution model developed by the Visual Geometry Group [42]. To enhance the learning process, VGG uses a small convolution filter (3x3), which increases the depth of the network [42], [43]. The VGG has three (3x3) convolutional layers, which are equivalent to having a single (7x7) convolutional layer. However, VGG uses three (3x3) convolutional layers to reduce the number of parameters as it contains more ReLU layers (one after each convolution layer), which makes the decision function more discriminatory [42], [44]. VGG has different variations depending on the number of layers used (VGG-11, VGG-13, VGG-16, and so on). The VGG-11 model consists of two stacks of convolution layers and a Max pool layer, followed by three stacks of two convolution layers and a Max pool layer, followed by three fully connected layers, resulting in having eight convolution layers and three fully connected layers [42], [45].

3. Residual Neural Network

ResNet is a deep neural network that was proposed by He *et al.* in 2015 for image detection [46]. In ResNet, the input of the layer is added to the output of the residual mapping, which can contain two or more layers. Fig. 2 shows that a shortcut connection is established between the input and output of the residual mapping along with an additional operation. This shortcut connection helps the network

learn more effectively, thereby improving its performance. ResNet is commonly used for image classification and object detection [47]. ResNet has different variations depending on the number of layers used (ResNet-18, ResNet-34, and so on). ResNet-18 comprises 17 convolution layers and a fully connected layer. A batch normalization layer and activation function can follow each convolution layer. The first convolution layer is followed by a max pooling layer. The network also includes eight sets of two convolutional layers, then an average pooling layer, and finally a fully connected layer with SoftMax activation function. The residual map is applied between the output of the even-numbered stack of convolution layers and the output of the next stack. The residual function is shown in (2):

$$y = F(x) + x \quad (2)$$

Where y represents the output vector of the layer, $F(x)$ represents the residual mapping to be learned, and x represents the input vector.

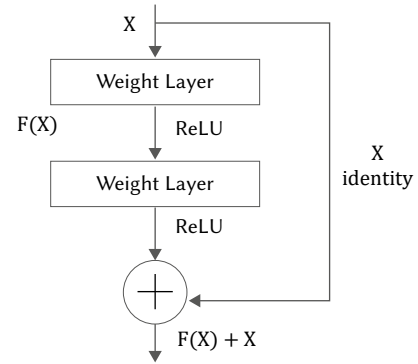


Fig. 2. Residual learning: a building block.

B. Experimental Setup

In this study, we evaluate the performance of the three learning models (CNN [2], ResNet-18 [46], and VGG-11 [42]), for image classification tasks using two datasets, Cifar-10 and Cifar-100 [48]. For ResNet-18, we replaced the batch normalization layers with group normalization as suggested and evaluated in [49]. The sizes of the three models transmitted by each client are presented in Table III. Cifar-10 dataset consists of 60,000 images that are categorized into 10 different classes. Each class has 6,000 images. Out of these 60,000 images, 50,000 are used for training and 10,000 for testing purposes. Similarly, the Cifar-100 dataset contains 60,000 images that have been classified into 100 distinct classes. Each class has 600 images. Out of these 60,000 images, 50,000 are used for training and 10,000 for testing purposes. In this study, we distributed the dataset among 100 clients in such a way that each client received 500 samples for training similar to [2], [22], [25]. We tried different settings and reported the best results obtained, we conducted the experiment over 300 rounds, with a learning rate of 0.01 and SGD as optimizer. At each round, we randomly selected 10 clients to participate as many studies used these settings [2], [25]. We performed the test on the client's side every five rounds. Following we show the experimental results of the three models, that highlight the model training accuracy and communication costs to reach a predefined target. We designed two cases to study the models' performance based on the data distribution. The first is the IID data distribution, and the second is the non-IID data distribution, following a similar setting as in [2]. In the IID data case, each client has data from all classes, where each client holds 50/5 samples from each class from the Cifar-10/Cifar-100 datasets. In the non-IID data setting, each client has data from a few classes, 2/20 classes, where each client holds 250/25 samples from the selected class from the Cifar-10/Cifar-100 datasets.

TABLE III. MODEL SIZE OF THE THREE MODELS TRANSMITTED BY EACH CLIENT

Dataset	CNN	ResNet-18	VGG-11
Cifar-10	8.22MB	41.61MB	104.87MB
Cifar-100	8.4MB	41.87MB	107.25MB

V. RESULTS

In this section, we present the experimental results of the three models with different numbers of epochs and batch sizes using two different datasets as we increase the computation per client. We compare the testing accuracy of the three models, communication bits exchanges and number of rounds to reach a predefined target accuracy.

A. Performance Comparison on Testing Accuracy

In this subsection, we show the performance of the three deep learning models using two cases when the clients have: (1) IID data and (2) non-IID data. Each case is evaluated using two different datasets, Cifar-10 and Cifar-100, and we varied the batch sizes and epoch values for each dataset.

1. IID Data Setting

To evaluate the performance of the three models, we tested them by varying the number of epochs and batch sizes. Fig. 3 illustrates the performance of the three models in the Cifar-10 and Cifar-100 dataset under the IID settings with an increase in batch size and local epoch value. In Cifar-10 the Fig. 3 (a) demonstrates that increasing the number of local epochs and decreasing batch sizes improves the model's performance for CNN model. Similar results were obtained in ResNet-18 model and VGG-11 model under the same settings, as shown in Fig. 3 (b) and Fig. 3 (c). The ResNet model converge faster with the increase of the local epoch value as shown in Fig. 3 (b), both batch sizes perform comparably well when trained with the same epochs value, indicating that the epoch count plays a crucial role in enhancing model performance. The VGG model in Fig. 3 (c) show a slow start especially with the 1-epoch configurations, however it obtains a higher accuracy with the increase of the global rounds with the 5-epoch configurations. In CNN and VGG-11 models, the best performance is achieved when the epoch size is 5, and the batch size is 16. While ResNet-18 performs best when the batch size =16 in all epoch values. We compared the performance of the three models at epoch=5 and batch size=16, 32, as shown in Fig. 3 (d). The VGG-11 model started slowly, but its performance improved with increasing rounds compared to ResNet-18 and CNN models. ResNet-18 and CNN provide comparable performance, as shown in Fig. 3 (d).

In Cifar-100 datasets, the performance of the three models under the IID settings is shown in Fig. 3. The models were evaluated with different numbers of epochs and batch sizes. Fig. 3 (e) shows the performance of the CNN model. The results indicate that the CNN shows better performance with a batch size of 16 in this setting. ResNet-18 and VGG-11 have similar performance, and they perform better with a batch size of 16 for different epoch sizes, as demonstrated in Fig. 3 (f) and Fig. 3 (g). The ResNet model converge faster with the increase of the local epoch value as shown in Fig. 3 (f), with the smaller batch size 16 outperforms the larger batch size 32. Also, the VGG model converge faster with the increase of the local epoch value as shown in Fig. 3 (g), with obtaining higher accuracy with the 5-epoch and 10-epoch configurations compared to the 1-epoch configurations for all batch sizes. We compared the performance of the three models at epoch=5 and batch size=16, 32, which is illustrated in Fig. 3 (h). The results indicate that VGG-11 performs worse compared to ResNet-18 and CNN models. Among these three models, CNN performs the best under the specified settings, as illustrated in Fig. 3 (h).

2. Non-IID Data Setting

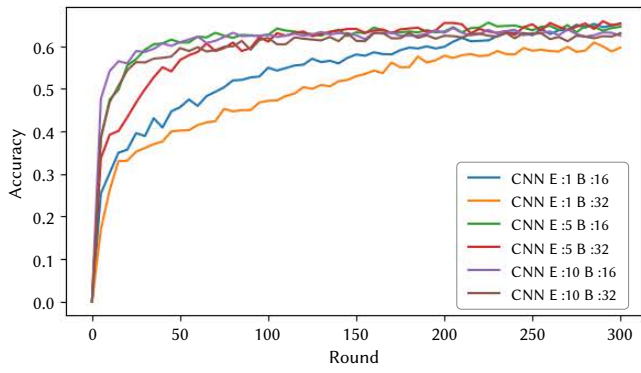
In the case of non-IID data settings, Fig. 4 illustrates the performance of the three models when tested using the Cifar-10 dataset and Cifar-100. In Cifar-10 the performance of all models improves as the number of local epochs increases and the batch size decreases as shown in Fig. 4 (a), Fig. 4 (b), and Fig. 4 (c). For all three model we observe that configurations with more epochs per round lead to higher accuracy, regardless of batch size. The best performance was achieved when the number of local epochs was equal to 5 for all models. However, the VGG-11 model performed the worst in non-IID settings compared to the CNN and ResNet-18 models, as illustrated in Fig. 4 (d). In Cifar-100 the performance of all models improves as the number of local epochs increases and the batch size decreases as shown in Fig. 4 (e), Fig. 4 (f), and Fig. 4 (g). For all three model we observe that configurations with more epochs per round lead to higher accuracy, regardless of batch size, with a slight edge for the smaller batch size 16. The best performance was achieved when the number of local epochs was set to 5 for CNN and VGG-11 models. However, in non-IID settings, CNN and ResNet with batch size = 16 perform better than the VGG-11 model that need more rounds to converge, as illustrated in Fig. 4 (h).

B. Performance Comparison on Communication Cost

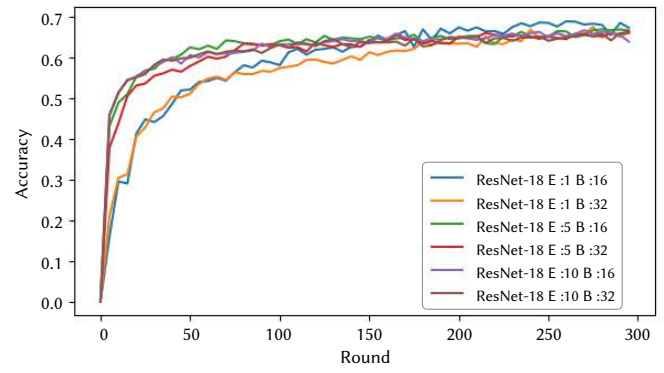
Communication cost is an important metric to evaluate FL, as the training process in FL is known to be a distributed process between clients, and requires clients to share their local models with a central server in multiple rounds through the network. For that, the number of rounds needed to reach a target accuracy and the number of bits sent by the clients are both an essential metric in FL. In this subsection, we evaluate the performance of the three learning models to achieve a predefined target accuracy in terms of the number of communication rounds (RoA@XX) and training bits exchanged from clients through the network. Table IV and Table V show the results for Cifar-10 and Cifar-100 datasets under the IID data settings, respectively. The “-” symbol means the target accuracy could not be obtained within the given number of communication rounds. However, the number of bits uploaded from clients during training is still reported.

Fig. 5 (a) illustrates the communication costs associated with different learning models for the CIFAR-10 and CIFAR-100 datasets. The figure shows that the CNN model consistently has the lowest communication cost across various configuration settings. In contrast, the VGG-11 model demonstrates the highest communication cost under all configurations. Notably, ResNet-18 falls between CNN and VGG-11 in terms of communication cost. Although ResNet-18 requires fewer communication rounds to achieve the target accuracy as shown in Table IV, and Table V, these rounds are more costly compared to those of the CNN model. This indicates that the CNN model can achieve the target accuracy with significantly lower communication overhead, making it a more efficient choice for federated learning scenarios.

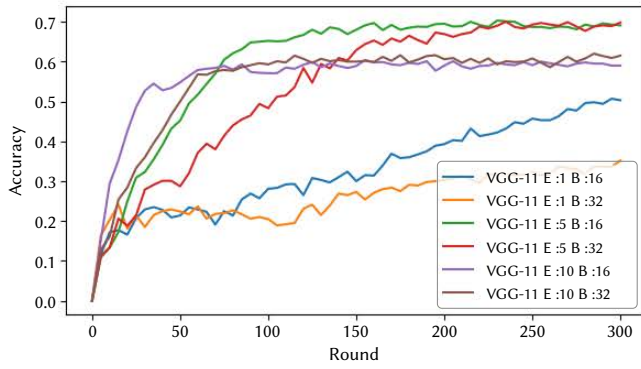
To sum up, the three models perform better when the data is IID data. In the Cifar-10 and Cifar-100 datasets, ResNet-18 shows faster convergence to reach the target accuracy compared to CNN and VGG-11 in most cases. However, since FL is a distributed learning process that shares a locally trained model instead of raw data to preserve privacy and provide an efficient communication, it is essential to consider the difference in the model weight of these three models. Although ResNet-18 requires fewer rounds, the number of bits transmitted is more than that of the CNN model as shown in Fig. 5.



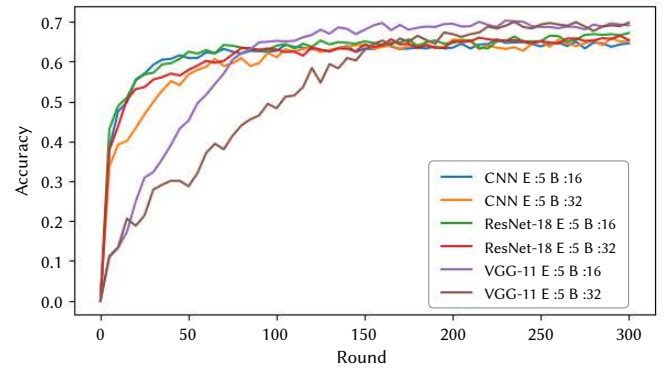
(a) CNN model for IID settings in Cifar-10



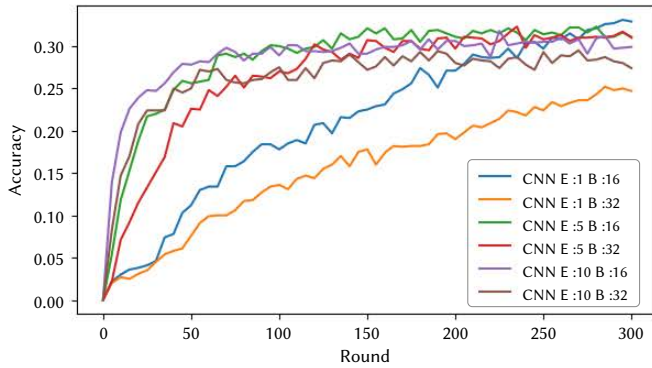
(b) ResNet-18 model for IID settings in Cifar-10



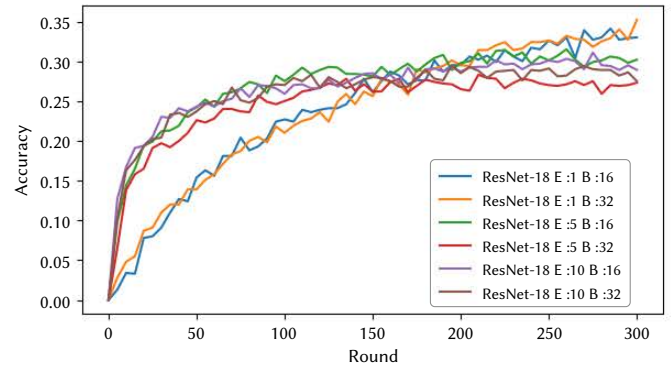
(c) VGG-11 model for IID settings in Cifar-10



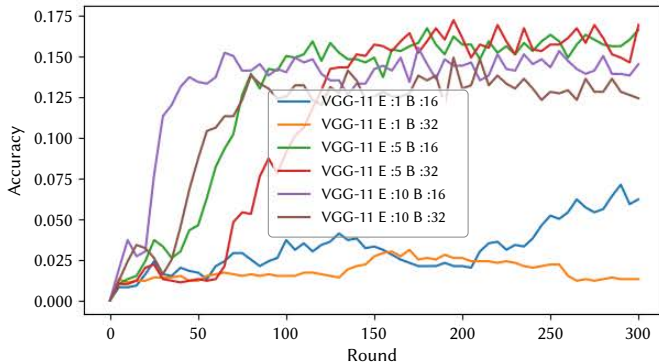
(d) The three models for IID settings in Cifar-10 with E=5



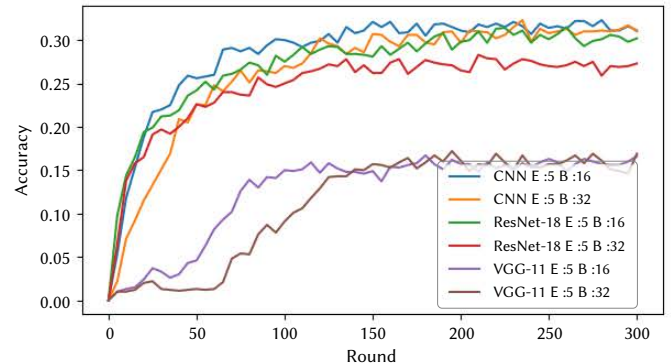
(e) CNN model for IID settings in Cifar-100



(f) ResNet-18 model for IID settings in Cifar-100

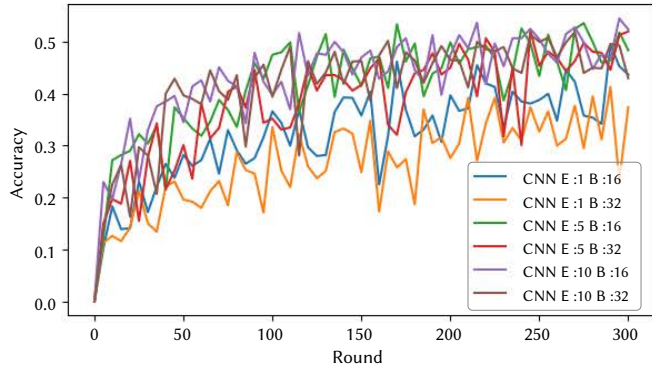


(g) VGG-11 model for IID settings in Cifar-100

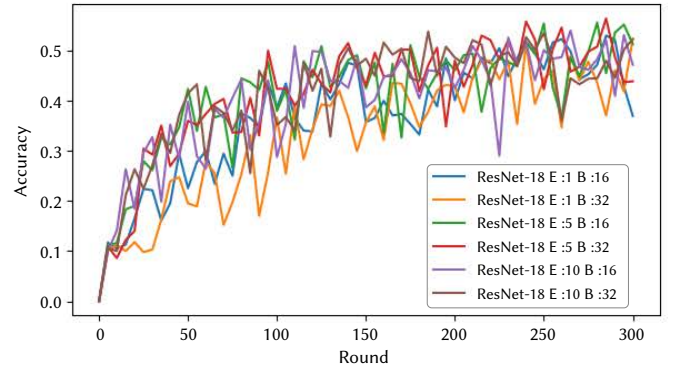


(h) The three models for IID settings in Cifar-100 with E=5

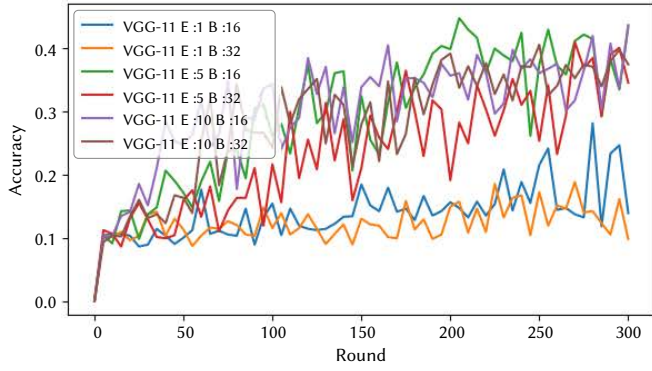
Fig. 3. The test accuracy of the three models for IID setting in Cifar-10 and Cifar-100 dataset.



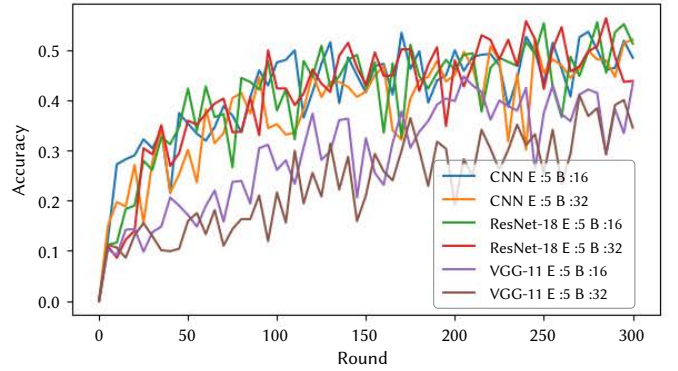
(a) CNN model for non-IID settings in Cifar-10



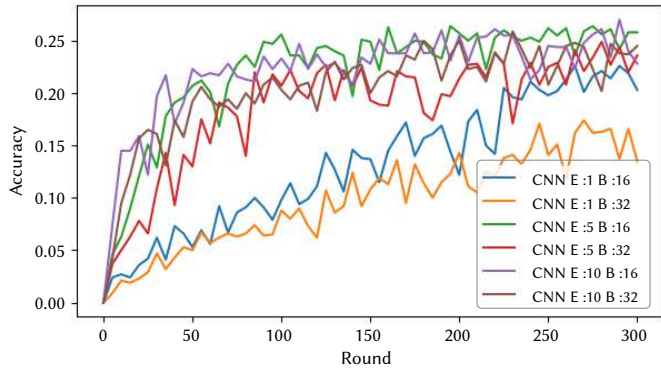
(b) ResNet-18 model for non-IID settings in Cifar-10



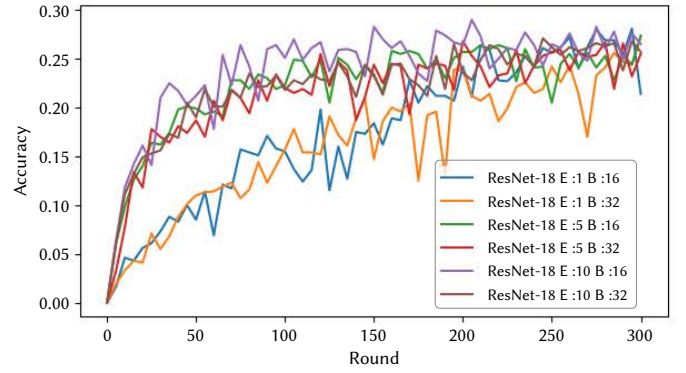
(c) VGG-11 model for non-IID settings in Cifar-10



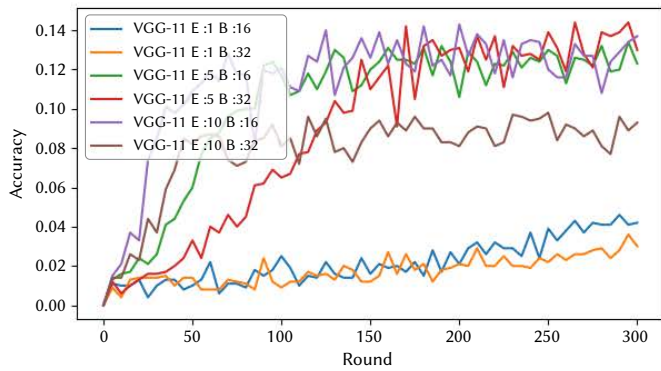
(d) The three models for non-IID settings in Cifar-10 with



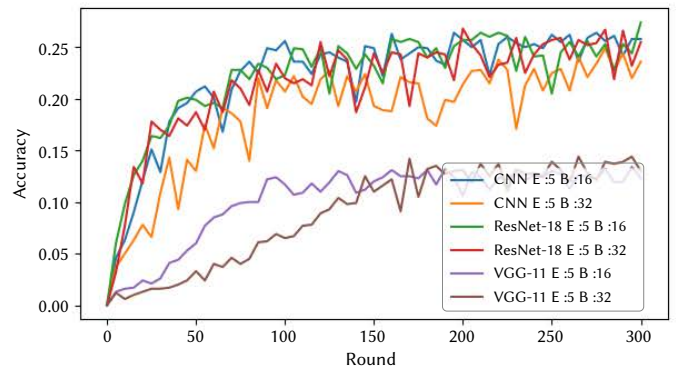
(e) CNN model for non-IID settings in Cifar-100



(f) ResNet-18 model for non-IID settings in Cifar-100



(g) VGG-11 model for non-IID settings in Cifar-100



(h) The three models for non-IID settings in Cifar-100 with E=5

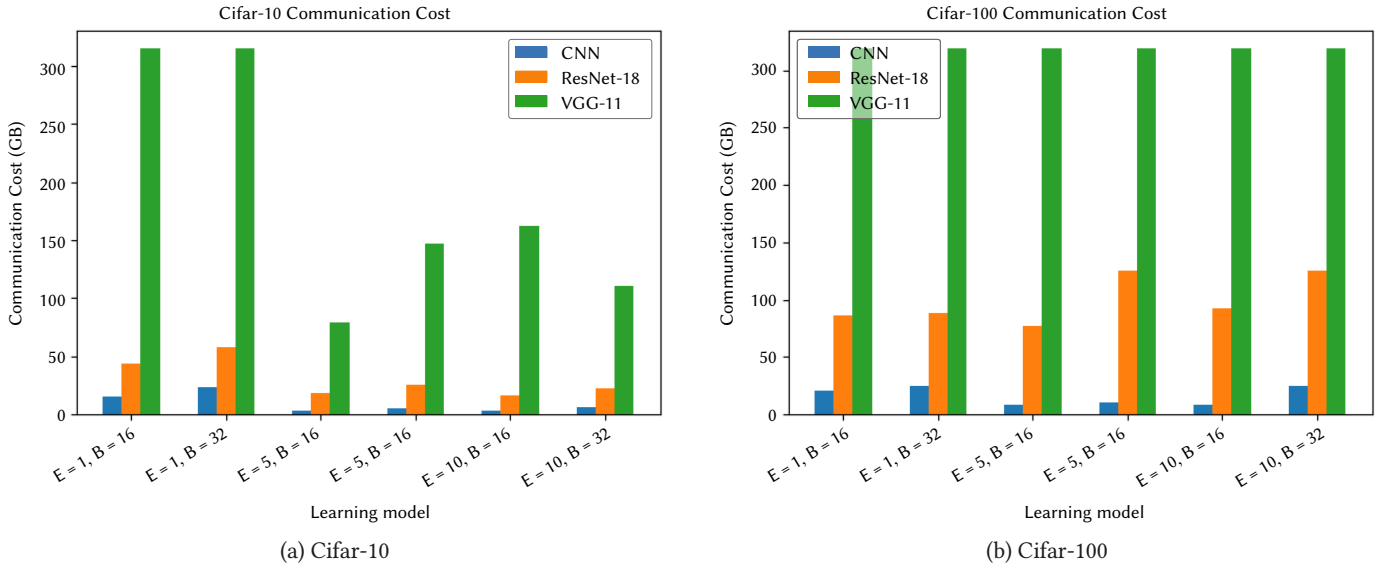
Fig. 4. The test accuracy of the three models for non-IID setting in Cifar-10 and Cifar-100 dataset.

TABLE IV. COMMUNICATION ROUNDS AND TRAINING BITS EXCHANGES TO REACH A TARGET ACCURACY FOR IID DATA SETTINGS IN CIFAR-10

Model	CNN						ResNet-18						VGG-11					
Epoch	1		5		10		1		5		10		1		5		10	
Batch	16	32	16	32	16	32	16	32	16	32	16	32	16	32	16	32	16	32
RoA@60	190	285	35	65	40	80	105	140	45	60	40	55	-	-	75	140	155	105
Communication Cost (GB)	15.63	23.44	2.87	5.34	3.29	6.58	43.70	58.27	18.73	24.97	16.65	22.89	314.62	314.62	78.65	146.82	162.55	110.11

TABLE V. COMMUNICATION ROUNDS AND TRAINING BITS EXCHANGES TO REACH A TARGET ACCURACY FOR IID DATA SETTINGS IN CIFAR-100

Model	CNN						ResNet-18						VGG-11					
Epoch	1		5		10		1		5		10		1		5		10	
Batch	16	32	16	32	16	32	16	32	16	32	16	32	16	32	16	32	16	32
RoA@30	240	-	95	120	105	-	205	210	185	-	220	-	-	-	-	-	-	-
Communication Cost (GB)	20.16	25.2	7.98	10.08	8.82	25.2	85.84	87.93	77.46	125.6	92.12	125.6	318.7	318.7	318.7	318.7	318.7	318.7

Fig. 5. The communication cost for the different models using Cifar-10/Cifar-100 dataset to reach target accuracy ($\approx 60\%$, $\approx 30\%$).

VI. DISCUSSION

In this study, we have assessed the performance of three different models, CNN, ResNet-18, and VGG-11, in image classification tasks under two different settings. Our findings suggest that VGG-11 has a slower start and necessitates more communication rounds to obtain the same accuracy as CNN and ResNet-18. Although VGG-11 eventually reaches a similar performance level to CNN and ResNet-18, it requires more communication rounds, and these rounds are costly as the VGG-11 model size is higher than CNN and ResNet-18. Moreover, VGG-11 did not perform well in the case of non-IID data setting. On the other hand, ResNet-18 performs well and converge quickly, requiring fewer communication rounds than CNN in some cases. However, it is noteworthy that ResNet-18 communication rounds cost more than CNN communication rounds, as the ResNet-18 model requires sending more bits when uploading the model.

The CNN model used in this study is a lighter model compared with ResNet-18 and VGG-11; however, it provides comparable performance compared to the other two models in terms of accuracy obtained in the predefined number of rounds. In some cases, more communication rounds may be required to obtain the same accuracy as ResNet-18; however, these rounds are less costly than ResNet-18 rounds since CNN requires fewer bits for exchanging the model compared to ResNet-18 and VGG-11. When training a model, obtaining high

accuracy is necessary and is considered an essential evaluation criterion. However, as FL is a decentralized approach that requires sharing clients' model with the central server, the communication cost is also considered a vital evaluation metric. Therefore, we must consider the number of rounds and the client bits sent to reach these results. When we use the communication costs as an evaluation metric to evaluate the performance of the three models, CNN provides better results than ResNet-18 and VGG-11. The CNN model exchanges fewer bits than ResNet-18 and VGG-11, providing comparable accuracy.

Based on our analysis of the performance of the three models and their associated communication costs, we recommend using a lighter model (such as CNN in [2]) in FL, since federated learning is a learning process that requires sharing locally trained models with a central server for several rounds through the network. Moreover, it is essential to determine the local epoch value since the devices in FL are limited in resources. Setting the local epoch to a high value does not indicate enhancing the model performance in these settings. Therefore, we recommend using an adaptive local epoch that can start with a higher value and decrease after a certain point to avoid overfitting and enhance the global model performance.

Hence, to answer the research question, Do we need a deeper network in a federated learning environment? Our answer is that in FL, it is necessary to consider different factors before choosing the

local model since the clients' devices are limited in resources and data, and the learning process is performed in rounds. We may not need a deeper neural network model since we need to consider not only model accuracy but also communication cost, and it may cost more to share a deeper model during training.

VII. CONCLUSION AND FUTURE WORK

Federated learning is a collaborative learning approach where the clients and server exchange training models for several rounds through the network to reach a predefined target. The choice of the learning model affects the model performance and the communication costs. Researchers have been faced with the decision of which learning model to choose when using FL for image classification tasks; several studies chose a deeper neural network model, such as VGG and ResNet, to evaluate their proposed work, while others chose a light model, such as CNN. In this study, we aimed to answer the question, "Do we need a deeper network in a federated learning environment?" Since FL is a decentralized approach, the model weight must be considered along with the model performance when choosing a neural network model since the model will be shared during training through the network. To answer this question, we conducted an empirical study investigating the impact of using three different neural networks in a FL environment (CNN, VGG-11, and ResNet-18). Our study evaluates the three models under different data settings (IID and non-IID) using two datasets (Cifar-10 and Cifar-100). We showed the performance of these models with varying numbers of local epochs and batch sizes. The results indicate that using CNN provides comparable results compared to the other models, with less communication cost; however, in some cases, it may require more rounds to reach the predefined target, but the communication cost (GB) is less than the other two models, making it a more practical choice for FL applications where communication efficiency is critical. We observed significant performance degradation for all models under non-IID settings compared to IID settings, highlighting the importance of addressing data heterogeneity in FL. Furthermore, our analysis revealed that using a 5-epoch configuration with a batch size 16 resulted in the best performance across all three models compared to other configurations. Our study provided valuable insights into the trade-offs between model accuracy and communication efficiency, suggesting that CNN offers a balanced approach by maintaining high performance while minimizing communication costs. Our findings indicate that training a model using a CNN model requires fewer network resources to train the FL model to reach accuracy similar to that obtained using deeper models.

For the future research direction, we aim there is a need to analyze the performance of FL on client device energy consumption and computational resources using different models and investigate if they are applicable on the client devices that are limited in resources. Also, investigate the effect of applying different compression method with deep neural networks. Furthermore, investigating the effects of introducing an adaptive local epoch size. By initially setting a higher epoch size that ultimately decreases after a certain point, to enhance the model's accuracy while simultaneously reducing costs.

VIII. FUNDING DECLARATION

1: The authors would like to acknowledge the support received from the Saudi Data and AI Authority (SDAIA) and King Fahd University of Petroleum and Minerals (KFUPM) under SDAIA-KFUPM Joint Research Center for Artificial Intelligence Grant no. JRC-AI-RFP-12.

2: This work has been partially supported by the project PCI2022-134990-2 (MARTINI) of the CHISTERA IV Cofund 2021 program;

by MCIN/AEI/10.13039/501100011033/ and European Union NextGenerationEU/PRTR for XAI-Disinfodemics (PLEC 2021-007681) grant, by European Commission under IBERIFIER Plus - Iberian Digital Media Observatory (DIGITAL-2023-DEPLOY- 04-EDMO-HUBS 101158511); and by EMIF managed by the Calouste Gulbenkian Foundation, in the project MuseAI.

REFERENCES

- [1] Z. Yang, M. Chen, K.-K. Wong, H. V. Poor and S. Cui, "Federated learning for 6G: Applications, challenges, and opportunities," *Engineering*, vol. 8, pp. 33-41, 2022.
- [2] B. McMahan, E. Moore, D. Ramage, S. Hampson and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, FL, USA, 2017.
- [3] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li and H. V. Poor, "Federated learning for internet of things: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 23, pp. 1622-1658, 2021.
- [4] X. Ma, J. Zhu, Z. Lin, S. Chen and Y. Qin, "A state-of-the-art survey on solving non-IID data in Federated Learning," *Future Generation Computer Systems*, vol. 135, pp. 244-258, 2022.
- [5] C. Carrascosa, F. Enguix, M. Rebollo and J. Rincon, "Consensus-based learning for MAS: definition, implementation and integration in IVEs," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, pp. 21-32, 2023.
- [6] M. Aledhari, R. Razzak, R. M. Parizi and F. Saeed, "Federated learning: A survey on enabling technologies, protocols, and applications," *IEEE Access*, vol. 8, pp. 140699-140725, 2020.
- [7] T. Li, A. K. Sahu, A. Talwalkar and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE signal processing magazine*, vol. 37, pp. 50-60, 2020.
- [8] S. AbdulRahman, H. Tout, H. Ould-Slimane, A. Mourad, C. Talhi and M. Guizani, "A survey on federated learning: The journey from centralized to distributed on-site learning and beyond," *IEEE Internet of Things Journal*, vol. 8, pp. 5476-5497, 2020.
- [9] C. Briggs, Z. Fan and P. Andras, "A review of privacy-preserving federated learning for the Internet-of-Things," *Federated Learning Systems: Towards Next-Generation AI*, pp. 21-50, 2021.
- [10] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie and R. Pedarsani, "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Online, 2020.
- [11] A. Khan, M. ten Thij and A. Wilbik, "Communication-Efficient Vertical Federated Learning," *Algorithms*, vol. 15, p. 273, 2022.
- [12] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li and Y. Gao, "A survey on federated learning," *Knowledge-Based Systems*, vol. 216, p. 106775, 2021.
- [13] B. Yu, W. Mao, Y. Lv, C. Zhang and Y. Xie, "A survey on federated learning in data mining," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 12, p. e1443, 2022.
- [14] L. Li, Y. Fan, M. Tse and K.-Y. Lin, "A review of applications in federated learning," *Computers & Industrial Engineering*, vol. 149, p. 106854, 2020.
- [15] Z. Lian, J. Cao, Y. Zuo, W. Liu and Z. Zhu, "AGQFL: Communication-efficient Federated Learning via Automatic Gradient Quantization in Edge Heterogeneous Systems," in *2021 IEEE 39th International Conference on Computer Design (ICCD)*, Storrs, CT, USA, 2021.
- [16] J. Xu, W. Du, Y. Jin, W. He and R. Cheng, "Ternary Compression for Communication-Efficient Federated Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, p. 1162-1176, 2022.
- [17] D. Rothchild, A. Panda, E. Ullah, N. Ivkin, I. Stoica, V. Braverman, J. Gonzalez and R. Arora, "Fetchsgd: Communication-efficient federated learning with sketching," in *Proceedings of the 37th International Conference on Machine Learning*, Virtual, 2020.
- [18] Y. Zhou, Q. Ye and J. Lv, "Communication-efficient federated learning with compensated overlap-fedavg," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, pp. 192-205, 2021.
- [19] Z. Qu, S. Guo, H. Wang, B. Ye, Y. Wang, A. Y. Zomaya and B. Tang,

- "Partial Synchronization to Accelerate Federated Learning Over Relay-Assisted Edge Networks," *IEEE Transactions on Mobile Computing*, vol. 21, pp. 4502-4516, 2021.
- [20] B. Alotaibi, F. A. Khan and S. Mahmood, "Communication Efficiency and Non-Independent and Identically Distributed Data Challenge in Federated Learning: A Systematic Mapping Study," *Applied Sciences*, vol. 14, p. 2720, 2024.
- [21] J. Zhong, Y. Wu, W. Ma, S. Deng and H. Zhou, "Optimizing Multi-Objective Federated Learning on Non-IID Data with Improved NSGA-III and Hierarchical Clustering," *Symmetry*, vol. 14, p. 1070, 2022.
- [22] X. Wu, X. Yao and C.-L. Wang, "FedSCR: Structure-based communication reduction for federated learning," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, pp. 1565-1577, 2020.
- [23] L. Gao, H. Fu, L. Li, Y. Chen, M. Xu and C.-Z. Xu, "Feddc: Federated learning with non-iid data via local drift decoupling and correction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, New Orleans, LA, USA, 2022.
- [24] Z. Lian, W. Liu, J. Cao, Z. Zhu and X. Zhou, "FedNorm: An Efficient Federated Learning Framework with Dual Heterogeneity Coexistence on Edge Intelligence Systems," in *2022 IEEE 40th International Conference on Computer Design (ICCD)*, Olympic Valley, CA, USA, 2022.
- [25] Y. Gong, Y. Li and N. M. Freris, "FedADMM: A robust federated deep learning framework with adaptivity to system heterogeneity," in *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, Kuala Lumpur, Malaysia, 2022.
- [26] S. Zhou, Y. Huo, S. Bao, B. Landman and A. Gokhale, "FedACA: An Adaptive Communication-Efficient Asynchronous Framework for Federated Learning," in *2022 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS)*, CA, USA, 2022.
- [27] X. Li, Z. Qu, B. Tang and Z. Lu, "Fedlga: Toward system-heterogeneity of federated learning via local gradient approximation," *IEEE Transactions on Cybernetics*, vol. 54, pp. 401 - 414, 2023.
- [28] Z. Chai, A. Ali, S. Zawad, S. Truex, A. Anwar, N. Baracaldo, Y. Zhou, H. Ludwig, F. Yan and Y. Cheng, "Tifl: A tier-based federated learning system," in *Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing*, Stockholm, 2020.
- [29] Q. Zeng, Y. Du, K. Huang and K. K. Leung, "Energy-efficient radio resource allocation for federated edge learning," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, Dublin, Ireland, 2020.
- [30] N. M. Jebreel, J. Domingo-Ferrer, D. S'anchez and A. Blanco-Justicia, "LFighter: Defending against the label-flipping attack in federated learning," *Neural Networks*, vol. 170, pp. 111-126, 2024.
- [31] H. Zhang, J. Jia, J. Chen, L. Lin and D. Wu, "A3fl: Adversarially adaptive backdoor attacks to federated learning," in *Advances in Neural Information Processing Systems*, New Orleans, LA, USA, 2024.
- [32] S. K. Lo, Q. Lu, C. Wang, H.-Y. Paik and L. Zhu, "A systematic literature review on federated machine learning: From a software engineering perspective," *ACM Computing Surveys (CSUR)*, vol. 54, pp. 1-39, 2021.
- [33] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278-2324, 1998.
- [34] B. Ghimire and D. B. Rawat, "Recent advances on federated learning for cybersecurity and cybersecurity for federated learning for internet of things," *IEEE Internet of Things Journal*, vol. 9, pp. 8229-8249, 2022.
- [35] Z. Lu, H. Pan, Y. Dai, X. Si and Y. Zhang, "Federated Learning With Non-IID Data: A Survey," *IEEE Internet of Things Journal*, pp. 1-1, 2024.
- [36] C. Janiesch, P. Zschech and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, pp. 685-695, 2021.
- [37] A. Mathew, P. Amudha and S. Sivakumari, "Deep Learning Techniques: An Overview," *Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2020*, pp. 599-608, 2021.
- [38] Z. Li, F. Liu, W. Yang, S. Peng and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, vol. 33, pp. 6999 - 7019, 2021.
- [39] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436-444, 2015.
- [40] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, L. Wang, G. Wang, J. Cai and T. Chen, "Recent advances in convolutional neural networks," *Pattern recognition*, vol. 77, pp. 354-377, 2018.
- [41] N. K. Chauhan and K. Singh, "A review on conventional machine learning vs deep learning," in *2018 International conference on computing, power and communication technologies (GUCON)*, Greater Noida, India, 2018.
- [42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [43] A. S. Rao, T. Nguyen, M. Palaniswami and T. Ngo, "Vision-based automated crack detection using convolutional neural networks for condition assessment of infrastructure," *Structural Health Monitoring*, vol. 20, pp. 2124-2142, 2021.
- [44] W. Wang, Y. Yang, X. Wang, W. Wang and J. Li, "Development of convolutional neural network and its application in image classification: a survey," *Optical Engineering*, vol. 58, pp. 040901-040901, 2019.
- [45] M. Pak and S. Kim, "A review of deep learning in image recognition," in *2017 4th international conference on computer applications and information processing technology (CAIPT)*, Kuta Bali, Indonesia, 2017.
- [46] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016.
- [47] M. Shafiq and Z. Gu, "Deep Residual Learning for Image Recognition: A Survey," *Applied Sciences*, vol. 12, p. 8972, 2022.
- [48] A. Krizhevsky, G. Hinton and others, "Learning multiple layers of features from tiny images," University of Tront, Toronto, ON, Canada, 2009.
- [49] K. Hsieh, A. Phanishayee, O. Mutlu and P. Gibbons, "The Non-IID Data Quagmire of Decentralized Machine Learning," in *Proceedings of the 37th International Conference on Machine Learning*, Virtual, 2020.

Basmah Alotaibi

Basmah Alotaibi received the B.Sc. degree in Computer Science from Imam Muhammad ibn Saud Islamic University (IMSIU) and the M.Sc. degree from the Department of Computer Science, King Saud University, Riyadh, Saudi Arabia. She is currently a Ph.D. student in Information and Computer Science, at King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia. Her research interest includes Federated learning, IoT, fog computing, cloud computing, and network security.



Fakhri Alam Khan

Fakhri Alam Khan is currently serving as an Associate Professor with the Department of Information and Computer Science at King Fahd University of Petroleum and Minerals. He is also a 'Research Fellow' with the Saudi Data and AI Authority (SDAIA) under the SDAIA-KFUPM Joint Research Center for Artificial Intelligence. He received his PhD in computer science from the University of Vienna, Austria, in 2010 and completed a post-doctorate from the Vienna University of Technology in 2017. He has published several research articles in various reputed peer-reviewed internationally recognized journals and has supervised numerous M.S. and Ph.D. students. His research interests include the IoT, data analytics, data provenance, distributed systems, machine learning, multimedia technologies, and nature-inspired metaheuristic algorithms.



Yousef Qawqzeh

Yousef Qawqzeh received the PhD. degree in systems engineering from UKM University, Bangi, Kuala Lumpur, Malaysia, in 2011, where he is currently working as an associate professor in the college of information technology, Fujairah University. He is currently working on several projects in the fields of machine learning, data science, and bioinformatics. He has several publications in international journal and conferences. His research interest includes the early prediction of cardiovascular diseases using the photoplethysmography technique, the development of computer-aided diagnosis systems for early diagnosis of breast cancer using artificial intelligence and machine learning techniques, and the detection and prediction of high-risk diabetics using machine learning and artificial intelligence techniques.



Gwanggil Jeon

Gwanggil Jeon received the B.S., M.S., and Ph.D. (summa cum laude) degrees from the Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea, in 2003, 2005, and 2008, respectively. From 2009.09 to 2011.08, he was with the School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada, as a Post-Doctoral Fellow. From 2011.09 to 2012.02, he was with the Graduate School of Science and Technology, Niigata University, Niigata, Japan, as an Assistant Professor. From 2014.12 to 2015.02 and 2015.06 to 2015.07, he was a Visiting Scholar at Centre de Mathématiques et Leurs Applications (CMLA), École Normale Supérieure Paris-Saclay (ENS-Cachan), France. From 2019 to 2020, he was a Prestigious Visiting Professor at Dipartimento di Informatica, Università degli Studi di Milano Statale, Italy. From 2019 to 2020 and 2023 to 2024, he was a Visiting Professor at Faculdade de Ciência da Computação, Universidade Federal de Uberlândia, Brasil. He is currently a professor at Incheon National University, Incheon. He was a general chair of IEEE SITIS 2023, and served as a workshop chairs in numerous conferences. Dr. Jeon is an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), Elsevier Sustainable Cities and Society, IEEE Access, Springer Real-Time Image Processing, Journal of System Architecture, and Wiley Expert Systems. Dr. Jeon was a recipient of the IEEE Chester Sall Award in 2007, ACM's Distinguished Speaker in 2022, the ETRI Journal Paper Award in 2008, and Industry-Academic Merit Award by Ministry of SMEs and Startups of Korea Minister in 2020.



David Camacho

David Camacho is full professor at Computer Systems Engineering Department of Universidad Politécnica de Madrid (UPM), and the head of the Applied Intelligence and Data Analysis research group (AIDA: <https://aida.etsisi.uam.es>) at UPM. He holds a Ph.D. in Computer Science from Universidad Carlos III de Madrid in 2001 with honors (best thesis award in Computer Science). He has published more than 300 journals, books, and conference papers. His research interests include Machine Learning (Clustering/Deep Learning), Computational Intelligence (Evolutionary Computation, Swarm Intelligence), Social Network Analysis, Fake News and Disinformation Analysis. He has participated/led more than 50 research projects (Spanish and European: H2020, DG Justice, ISFP, and Erasmus+), related to the design and application of artificial intelligence methods for data mining and optimization for problems emerging in industrial scenarios, aeronautics, aerospace engineering, cybercrime/cyber intelligence, social networks applications, or video games among others. He serves as Editor in Chief of Wiley's Expert Systems from 2023, and sits on the Editorial Board of several journals including Information Fusion, IEEE Transactions on Emerging Topics in Computational Intelligence (IEEE TETCI), Human-centric Computing and Information Sciences (HCIS), and Cognitive Computation among others.

Measuring the Difference Between Pictures From Controlled and Uncontrolled Sources to Promote a Destination. A Deep Learning Approach

Angel Diaz-Pacheco¹, Miguel A. Álvarez-Carmona^{2,6}, Ansel Y. Rodríguez-González^{3,6}, Hugo Carlos^{4,6}, Ramón Aranda^{5,6*}

¹ Universidad de Guanajuato - Departamento de Ingeniería Electrónica - Campus Irapuato-Salamanca, Salamanca (Mexico)

² Centro de Investigación en Matemáticas - Unidad Monterrey, Apodaca (Mexico)

³ Centro de Investigación Científica y de Educación Superior de Ensenada - Unidad de Transferencia Tecnológica Tepic, Tepic (Mexico)

⁴ Centro de Investigación en Ciencias de Información Geoespacial - Unidad Mérida, Mérida (Mexico)

⁵ Centro de Investigación en Matemáticas - Unidad Mérida, Mérida (Mexico)

⁶ Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCyT), Ciudad de México (Mexico)

* Corresponding author. arac@cimat.mx

Received 14 January 2023 | Accepted 28 September 2023 | Early Access 30 October 2023



ABSTRACT

Promoting a destination is a major task for Destination Marketing Organizations (DMOs). Although DMOs control, to some extent, the information presented to travelers (controlled sources), there are other different sources of information (uncontrolled sources) that could project an unfavorable image of the destination. Measuring differences between information sources would help design strategies to mitigate negative factors. In this way, we propose a deep learning-based approach to automatically measure the changes between images from controlled and uncontrolled information sources. Our approach exempts experts from the time-consuming task of assessing enormous quantities of pictures to track changes. To our best knowledge, this work is the first work that focuses on this issue using technological paradigms. Notwithstanding this, our approach paves novel pathways to acquire strategic insights that can be harnessed for the augmentation of destination development, the refinement of recommendation systems, the analysis of online travel reviews, and myriad other pertinent domains.

KEYWORDS

Destination Image, Deep Learning, Destination Marketing Organization, Scene Recognition, Natural Language Process.

DOI: 10.9781/ijimai.2023.10.003

I. INTRODUCTION

THE World Tourism Organization reported that in 2019 the travel and tourism industry generated approximately \$1.5 trillion (USD) and was the third-largest export category in the world [1]. This industry is of great importance to many countries around the world that devote considerable resources to the research and promotion of their attractions. It is also important to stress that the paradigms of the tourism sector have changed with the massification of the Internet and social networks. The high availability of information and social media have promoted the surge of new dynamics in tourism [2]–[4]. *Big Social Data* and User Generated Content (UGC) are becoming key sources of well-timed and rich knowledge, supporting data-driven decision approaches that address the management of complex relationships through the use of emerging and comforting technology [5]. Online information on

climate, transport, attractions, and accommodation offers to tourists a starting point for planning their holidays, whose decisions are strongly influenced by the opinions of other tourists shared electronically [6].

One of the main activities of Destination Marketing Organizations (DMOs) is the implementation of marketing campaigns to promote tourism and to design strategies to deal with adverse circumstances affecting the industry [7],[8]. These situations are also present on the internet in the form of information coming from uncontrolled sources, which sometimes deteriorates the projected *destination image*. Monitoring changes in photographs that differ greatly from the projected *destination image* by the DMOs could help to design procedures to mitigate these effects. However, tracking these changes imposes certain limitations, such as the availability of experts and the manpower to overcome the enormous task of analyzing the huge number of photographs daily posted on the internet.

Please cite this article in press as: A. Diaz-Pacheco, M. A. Álvarez-Carmona, A. Y. Rodríguez-González, H. Carlos, R. Aranda. Measuring the Difference Between Pictures from Controlled and Uncontrolled Sources to Promote a Destination. A Deep Learning Approach, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 4, pp. 18-31, 2025, <http://dx.doi.org/10.9781/ijimai.2023.10.003>

In this work, we propose an approach based on deep learning techniques to automatically measure changes between the photographs used to promote a destination by the DMOs (controlled sources) and those published on the Internet (uncontrolled sources) during adverse events that negatively could influence the projected destination image. To our best knowledge, this is the first effort to propose a method to deal with this problem, which would represent new avenues for collecting and refining strategic information and potential applications for tourists or DMOs.

II. DESTINATION IMAGE IN TOURISM

The term *Destination Image* in Tourism or DIT has been around since the 1970s. DIT is the total impression of the destination in the minds of tourists and residents [9]. In a broader sense, the DIT is related to a set of ideas, impressions, and beliefs that a person or group shares about a particular place based on long-term information obtained from various sources that led to the construction of a positive/negative image [10], [11]. Several studies have extensively investigated the process of formation of the destination image. Among these, Gunn [12] influenced and contributed to identifying the levels of image formation based on the type of information available to the tourist. This framework proposes three main levels: organic, induced, and modified-induced. The organic level refers to general ideas that a person has about a particular place; this information can be obtained from multiple sources, such as personal conversations and television. The induced image is formed using information received and processed intentionally by the tourism industry, including brochures and advertisements produced by DMOs. As for the modified-induced image, it refers to the mental reconfiguration of the DIT derived from the travel experience within the destination. These levels are incorporated in the seven steps presented in Fig. 1. According to González-Rodríguez et al. [13], the choice of a particular destination is influenced by a more positive and stronger destination image, which is shaped by different secondary factors such as media on the internet [14], [15]; Frías et al. [16] stressed the importance of the impact of pictures of a destination, because the processing of an image requires less cognitive resources and affects all users; As for social media, the studies of Gallarza et al. [17] and Govers and Go [18] suggested that in the presence of paradigms such Big Data, the destination image perceived by tourists would be influenced by more diverse sources of information, which makes it difficult to quantify this construct because it tends to be complex, relativistic and dynamic.

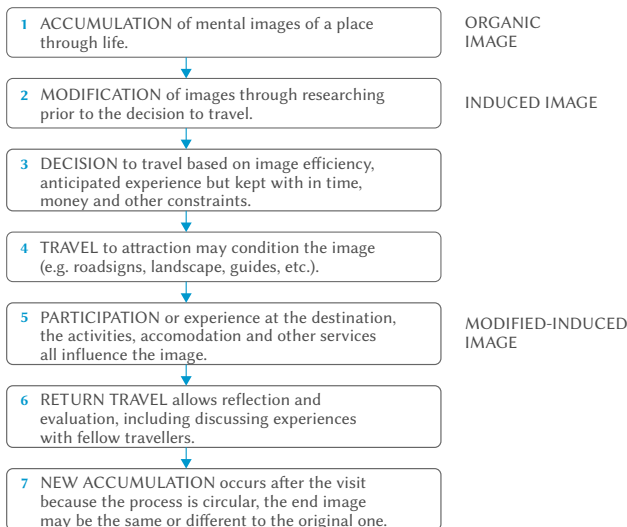


Fig. 1. Stages in the process of DIT formation. Source: [12].

The DIT construct can be classified into different categories. The *projected image* can be defined as the ideas and impressions about a place that is available for the consideration of potential tourists. These secondary images may come from induced sources for the sole purpose of promoting the destination to attract tourists or from autonomous sources (news, travel magazines, social networks, ads, etc.) which are not controlled by DMOs that aim to supply information for travelers [19]. On the other hand, the *perceived images* come from organic sources of two main types: people who talk with their friends and family and UGC on different electronic media. Perceived images are the product of interaction between projected images and the characteristics of potential visitors, but the most credible source is the personal experience of previous visits to the tourist destination [20].

Given the importance of factors such as social networks and the Internet in the image formation, monitoring the variations between the image projected by DMOs and images on the Internet could allow the design of emergency strategies to mitigate the effects of this bad publicity on potential tourists. With new paradigms like Big Data and deep learning, the analysis of different aspects of the destination image in tourism (DIT) construct has been automated and applied to large amounts of information. In our research, we have established a classification for studies. Those who have analyzed DIT's construct using computer techniques can be divided into the following categories: Studies that are aimed at identifying unique attributes of DIT using User Generated Content to get ideas for destination branding [21], [22]; Studies exploring the evolution of the construct using content analysis, such as changes in image perceived during economic crises [23]–[25]; Studies using social media and big data to analyze the construct [26]–[28]; Research interested in understanding the underlying factors influencing the construction [29], [30]; The research focused on the use of multimedia content to achieve better strategies for DMOs [31], [32]; Studies aimed at exploring the factors in tourist attractions to understand the choices of tourists [33], [34]; Finally, another studies do not fall into one category [35], [36]. Based on the many important approaches in the literature, below, we summarize the most influential in our research. The following studies propose methods for analyzing images (photographs):

- Leung et al. [37] carried out a three-stage analysis. In the first, they got pictures of flicker. The second stage tried to find popular places using a clustering method to differentiate images from different users and with certain proximity based on the coordinates of photos. The idea behind this logic was that many tourist photos in the same place mean that such a place is popular. With this data, the authors tried to analyze the patterns of tourist visits and to identify the image of the destination through the analysis of these pictures.
- Nixon [38] proposed a machine learning method to segment the audience of visual destination imagery in order to reflect that different types of audiences will build divergent visual destination images. He used the MachineBox.io service to train a visual classifier for image annotation and to obtain the frequency of these classes. Comparisons between different destination images were made (a ground truth of manually annotated images in comparison to automatic approaches).
- Xiao et al. [39] proposed a method for obtaining a seasonal variation index of the major tourist scenes in the province of Jiangxi in China. The 22 main scenes were analyzed using the Deep learning model Places365 to obtain representative labels of the destination photos, and they applied the Linear Discriminant Analysis (LDA) algorithm to get a distribution of the major themes in those scenes.
- Arabadzhyan et al. [40] presented a method for measuring the dissimilarity among destination images of different places and how such images are influenced by endogenous and exogenous factors. They used the Google Cloud Vision API for image labeling,

and distances were calculated as the absolute value of the average distances of the top K labels.

- Bui et al. [41], proposed a big data framework for measuring the DIT construct using machine learning techniques. The framework was able to process textual and visual data, separating it into sub-constructs, seasons, and tourist characteristics. The measure was carried out in each sub-construct (Nature and landscape, people and culture, history, among others) through the frequency of words and sentiment analysis of all information.
- He et al. [42] explored how to analyze UGC to obtain the characteristics in a core-periphery scheme of the destination image to achieve a better selection of images for DMO's to promote the destination. The DeepSentiBank algorithm was used to detect emotions and objects from photos of three different data sets (two of Organization Generated Content and one of UGC). Labels have been processed and used to build a semantic network (with the algorithm NetworkX) for representing the structure of the perceived Destination Image. With this network, the centrality scores were computed to generate a list of major adjectives and nouns ranked. With these metrics as a goodness measure, the best DMO's photos were chosen, and a regression algorithm was used to evaluate their online engagement.

Concerning pertinent studies that have undertaken analogous tasks from a technological perspective and have exerted a notable influence on the current work, a delineation is provided subsequently. The above studies have researched different aspects of the DIT construct and sought marketing information by using tourist segmentation, analysis of popular attractions, and other techniques. The present study aims to measure the discrepancy between photos from controlled sources (DMOs) and those published online for newspapers, magazines, and users (uncontrolled sources) during adverse events that can negatively impact the pre-visit destination image. On considering the above, the following questions have been proposed to guide our research.

- How can images be used to measure differences in the projected destination image between controlled and uncontrolled sources?
- Is our method capable of pointing critical points at measured differences?
- Is our method able to measure those differences for different destinations?

From the analysis of the related work, we have obtained valuable information utilizing the convenience of deep learning and the use of Big Data to derive better tourism intelligence and new potential applications using massive amounts of multimodal data. All these studies influenced our work and the proposed approach is presented in subsequent sections.

III. PROPOSED APPROACH

We propose a methodology to automate the process of measuring changes between the photographs used to promote a destination by the DMOs (controlled sources) and those published on the internet (uncontrolled sources) during events that negatively could influence the projected *destination image*. To this aim, we perform a comparative image analysis on three different periods: an analysis of previous the event, an analysis of when the event took place, and a last analysis of when the event has finished. In this way, we can measure the behavior of the uncontrolled sources in comparison to controlled sources.

To automatically perform the image (photographs) analysis, we used a collection of tools and methods based on three stages: scene recognition, characterization of features (feature management), and measurement of the difference between features. Each stage is explained in detail below.

A. Scene Recognition

Scene recognition provides a description of the content of the image rather than listing objects in the scene. The main objective of this task is to assign semantic labels to images and, such labels are defined by real observers to equip computers with the ability to describe scenes like humans [43]. The task of object recognition, on the other hand, is postulated as a representation of the object being investigated with the image features available while rejecting the background features [44]. From the above, we can see that scene recognition provides richer descriptions not only of the objects in the scene but also of their surroundings and context. For these reasons, we opt for using the scene recognition approach. Thus, we choose the well-known Places-CNNs [45], which is a deep learning tool formed by a set of convolutional neural networks. Places-CNN uses different CNN architectures, but according to their results, the most precise are VGG-16 and ResNet152, which likely are at the core of the implementation we used. A typical output of Places-CNN is a set of tags that describe the scene and objects in it. For example, the photograph of a food area in a shopping mall (Fig. 2) is described by Places-CNNs with the terms and strengths in Table I. Tags in *Categories* are the elements discovered on the scene with the deep learning modules. We have named strength to the degree of confidence a neural network has in declaring that a particular element is on the scene. Scene attributes are general descriptions of the image analyzed, such as *closed area*, which means that the action takes place inside. Taking into account the above information, we can imagine Places-CNNs as an advanced device for analyzing huge data loads, recognizing elements in scenes (*attributes*), and with a degree of certainty (*strength*) in this detection process.



Fig. 2. Scene analyzed with Places-CNNs.

TABLE I. INFORMATION OBTAINED BY PLACES-CNNs BY ANALYZING THE SCENE IN FIG. 2

	Scene Recognition	Strength
Categories	Food_court	0.511
	fastfood_restaurant	0.085
	cafeteria	0.083
	dining_hall	0.040
	flea_market/indoor	0.021
Scene attributes	No horizon	0.903
	Enclosed area	0.495
	Man-made	0.444
	Socializing	0.423
	Indoor lighting	0.211
	cloth	0.151
	Working	0.102
	Congregating	0.081

B. Characterization of the Feature Space

Once a photo is processed by Places-CNN, we use as *photo-features* only the outputs given by the attributes of the scene because, during the experimental phase of this work, we have experienced significant variations in the use of both sets of features (categories and attributes). Therefore, we consider that a more stable performance can be achieved using only *attributes*. That is because they are more descriptive characteristics in a wider scope.

Given a set of features from a set of images, we characterize them as a histogram, where the bins are the label of the features (attributes of the scene), and the sum of the corresponding strength as the *frequency*. Thus, we compute a total of four histograms: one for the set of pictures obtained by the DMOs from a specific destination (controlled photographs), and one for the event on each period to analyze (pre-event, during-event, and post-event), from uncontrolled sources. It is important to note that, the set of photographs does not necessarily have the same number of pictures or features, hence, we normalize the histograms to keep equal scales. In the following section, we briefly explain a method that can be used to quantify the differences between histograms.

C. Measurement of Difference Between Features

To measure the differences between the computed histograms, we use the Earth Mover's Distance (EMD). EMD is a distance metric that measures the dissimilarity of two histograms. The computer vision community has been enthusiastic with this technique [46]. Let's consider two normalized histograms $q = (q_1, \dots, q_n)$ and $p = (p_1, \dots, p_m)$, each with n and m bins, respectively; q_i and p_j are frequency values of the histograms p and q , for the bin labels i and j , respectively. F is a *flow matrix*, where f_{ij} is the flow to move from q_i to p_j , and a *cost matrix* C , where c_{ij} means the cost of moving flow from the bin label i -th of q to the bin label j -th of p . The total cost of moving the unit flow to F and C between the histograms q and p can be defined as:

$$d(q, p) = \sum_{i=1}^n \sum_{j=1}^m f_{i,j} c_{i,j} \quad (1)$$

The ground distance (here called cost matrix) C can be designed, depending on the behind problem, by experts in the field or derived from a formula. It is clear that $c_{ij} = 0$ and a greater distance between label bins i and j means a greater $c_{ij} > 0$. As p and q are normalized, then $\sum_{i=1}^n q_i = \sum_{j=1}^m p_j = 1$, and the EMD between p and q is defined as follows:

$$\begin{aligned} \text{EMD}(q, p) = \min_F d(q, p), \\ \text{subject to} \\ f_{i,j} \geq 0, \forall i \in [1, n] \text{ and } \forall j \in [1, m], \\ \sum_{j=1}^m f_{i,j} = q_i, \forall i \in [1, n], \\ \sum_{i=1}^n f_{i,j} = p_j, \forall j \in [1, m], \\ \sum_{i=1}^n q_i = \sum_{j=1}^m p_j = 1. \end{aligned} \quad (2)$$

The idea behind the equation (2) is to find the minimum cost to transform q to p . The computed cost represents the difference between histograms, it is to say, if p and q are equal then $\text{EMD}(q, p) = 0$; otherwise $\text{EMD}(q, p) > 0$.

Since a set of photographs is characterized as a histogram, two sets of different photographs can be compared by solving (2). However, a fundamental aspect to consider is the ground distance (or cost matrix C) among each bin label in the histogram. In the problem under study, each bin label of our histograms is a scene attribute, thus obtaining the distance between attribute concepts is not a trivial task. The following section discusses three different approaches to addressing this issue.

1. Ground Distance

As indicated in the previous section, a ground distance matrix must be created to measure the differences between two histograms. When the histograms are about specific entities such as colors, we can use different analytical methods to calculate the differences among them; however, if the histogram is for items like different kinds of wines, an expert or a group of them is needed to provide a measure about how different are a "Pinot noir" and a "Merlot". In the problem at hand, we need to provide a measure of the semantic distance between concepts such as "man-made" and "indoor light", we opted to use two well-known approaches in the field of deep learning and one last rather straightforward. Since there are no available comparisons of the performance of those deep learning approaches for tasks like the one in this study, we have decided to use all of them to observe their suitability for our method. In the following, we explain three different proposals to compute the ground distance C .

WordNet The first proposal to compute Ground distance is based on Natural Language Processing (NLP). This approach is performed by using the WordNet database. WordNet is a large lexical database of English. It contains almost 80,000 noun word forms organized into 60,000 lexicalized concepts. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked utilizing conceptual-semantic and lexical relations [47]. To measure the distance of two concepts, we used the Wu & Palmer similarity measure (*wup*), which calculates the relatedness by considering the depths of two synsets in the WordNet taxonomies, along with the depth of the Least Common Subsumer (LCS) [48]. The LCS of two concepts A and B is the most specific concept, which is an ancestor of both A and B [49]. The calculation of the metric is given by:

$$wup = 2 * \frac{\text{depth}(\text{lcs}(s_1, s_2))}{\text{depth}(s_1) + \text{depth}(s_2)} \quad (3)$$

where s_1 and s_2 are the synsets to be compared, depth is the deepness in the taxonomic tree and lcs refers to the Least Common Subsumer.

Word embeddings The second proposal, also based on the NLP approach, was a technique known as *word embeddings*. This technique performs a mapping of a categorical variable (a word) to a vector of continuous numbers. For implementation, we use the well-known Word2Vec method [50]. One of the major advantages of these models is that the distributed representation reaches a level of generalization that is not possible with classic models [50]. A typical n-gram model operates in terms of discrete units that have no intrinsic relationship between them, a continuous model works in terms of word vectors, therefore, similar word have similar vectors. For our experiments, we use the *GoogleNews-vectors-negative300* pre-trained model. This model has been trained on a part of the Google News data set and contains 300-dimensional vectors for 3 million words and phrases [51]. By using a vector representation for each word that preserves semantic meaning, we can approach the quantification of similarity as a simple Euclidean distance between the points in a 300-dimensional space.

Constant Value Finally, for comparison purposes, the last approach to compute Ground distance is assuming that the distance between all elements is constant. This means that the difference between all *scene attributes* is the same. Specifically, we set the value constant equal to 1. For example, the distance between *indoor* and *eating* is 1 and it is the same for *eating* and *man-made*.

D. Summary

Our approach follows a three-stage process: obtain the *attributes of the scene* from a set of pictures, characterize output attributes into normalized histograms, and compute the differences between the histograms (see Fig. 3 and Fig. 4). In the *first stage*, a set of photos

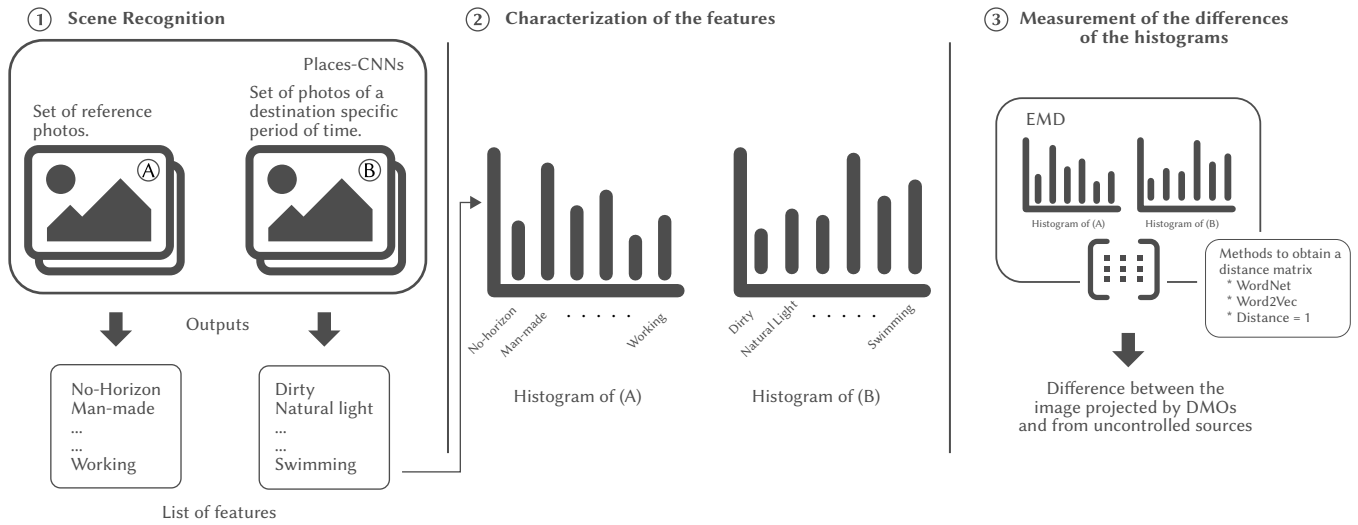


Fig. 3. Graphical description of the phases of the proposed approach.

is processed using the Places-CNN's tool for scene recognition. The result in each picture is a set of attributes and strengths (confidence level of the neural network) of the scene. In the *second stage* the complete set outputs are characterized on a histogram, where the bins are the label of the features (attributes of the scene), and the sum of the corresponding strength as the "frequency". Since the strengths are continuous values and the set of features are not necessarily the same we normalize the histograms to keep equal scales. In order to have a baseline to compare, we built four histograms: from a reference set of a collection of photos of the tourist place, and three from a set of photos (of the same reference place) taken on the internet on different periods of time around an event (pre-event, during event and post-event). The *third step* is to calculate the differences between the histogram obtained on the reference photos of the tourist destination and the histogram from the three periods under evaluation. In this stage, the most important thing is the adequate generation of the distance matrices, which is necessary to calculate the difference between histograms. The result of this step is a symbolic measure of the differences. With the elements in the histograms, we can also perform set operations to know more about each period analyzed.

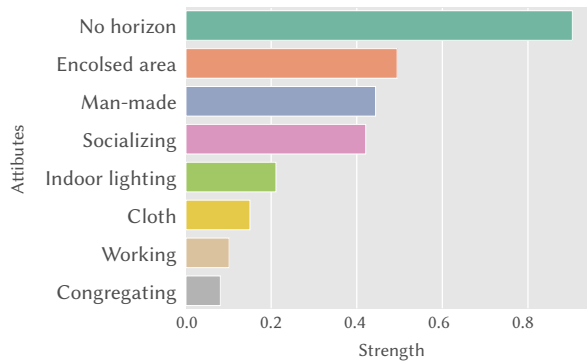


Fig. 4. Example of a histogram of the attributes of Fig. 2.

IV. EXPERIMENTAL DESIGN

A common practice for travelers from all over the world is to search the Internet for details of where they are about to visit, especially photos. Considering the above, we want to know how the pre-visit destination image is affected by pictures of adverse events available on the internet as climatic or social phenomena. To verify whether the changes can be automatically quantified from photographs, we select a

set of events with extensive media coverage that took place in Mexico. We employed an approach based on periods, labeled as "during" for the time the event took place and "before" and "after" for the pre- and post-event times. The photos were searched and downloaded manually, and equal proportions of photos were gathered for all time periods. For the collection of images, we implemented a web scraper. This algorithm automates the search and downloads the pictures. We have used Google Images, but any web platform may be employed. The terms used for the search follow the structure shown below:

"City" "streets" "Larger than 800x600" "after:Date_1 before:Date_2"
 "City" "beaches" "Larger than 800x600" "after:Date_1 before:Date_2"
 "City" "news" "Larger than 800x600" "after:Date_1 before:Date_2"

The term "beaches" is used in the sun and beach destinations. The deep learning module needs images of at least "224x224" pixels, so the size can be introduced to the search to avoid smaller images. We used both English and Spanish for the terms. The events are described in Section IV.A. To make comparisons, we also need a baseline of the places where the events took place. We used a set of DMO's and Flickr's photographs to build it (see Section IV.B). The results are presented in Section V.

A. Events Under Analysis

To test our method, we select a small sample of adverse events that occurred in different destinations in Mexico. To be able to ensure sufficient content, we choose events with extensive media coverage. We reduced our list only to events that affected tourist destinations and these are Mexico City, Guadalajara, Cancun, Acapulco, and San Blas. Apart from Mexico City and Guadalajara, the others are well-known sun and beach destinations for foreign (Cancun, Acapulco) and domestic (San Blas) tourists. Mexico City and Guadalajara are recognized for their historical relevance and comprehensive range of tourism-related services. Mexico City was recognized by UNESCO as a World Heritage Site in 1987 and the Museo Cabañas in Guadalajara and the surrounding Agave landscape got such recognition in 1997 and 2006 respectively. In addition to its beautiful turquoise beaches, Cancun is known for its World Heritage Maya ruins. In 2019, it was visited by 463,428,131 tourists, being the second most-visited destination behind Mexico City. Acapulco is a tourist attraction in the Mexican state of Guerrero, located on the Pacific coast of Mexico. With an annual average temperature of 28° Celsius, Acapulco is one of the few destinations in the world with year-round pleasant weather. It offers a variety of services to enhance the visiting experience in addition to its natural charms. In 2019 it was the third most popular

destination in Mexico with 186,377,996 visitors. San Blas is part of a new tourism development called Riviera Nayarit, which promises to bring tourism development to the coastal region of the municipalities of the region, an area rich in diverse natural and cultural resources. In 2019, above 278,513 tourists visited San Blas. For the occupation data, we have relied on statistical information gathered by the governmental organism in charge of tourism in Mexico, the Secretary of Tourism or SECTUR under the acronym in Spanish [52].

1. Sargassum Season in Cancun 2021

Sargassum is a species of algae that is abundant in tropical zones like the Caribbean coasts [53]. Since 2019, there has been a considerable increase in the proliferation of Sargassum, but the highest concentration occurs during the hottest months of the year. To characterize this event, we selected the photographs posted between March and June 2021. The images for the previous (**before**) period were from January and February of the same year. For the following period (**after**), the months of July and August were taken into account. Fig. 5 presents some thumbnails of representative images of the event. 100 images were collected for each time period for this event. This number of images was chosen to avoid repetitions among the photos and to ensure diversity.



Fig. 5. Representative images of the season of sargassum in Cancun.

2. Protesters in Mexico City During 2021

Before describing this event, we want to state that we are not against the protesters and their struggle for women's rights. Our interest is only from the point of view of tourism to analyze the impact of this social phenomenon on the sector. On March 8, International Women's Day, police and activists clashed in Mexico City during a march [54]. According to the BBC news portal, officers pushed back the protesters with tear gas and riot shields. For the event, we collected photographs from 03-08-2021 to 03-10-2021. The before period was from 02-26-2021 to 02-28-2021. For the after period, the dates range from 04-10-2021 to 04-12-2021. Fig. 6 shows some representative images of the event. 100 images were collected for each time period for this event.

3. The Effects of Hurricane Willa on San Blas, Nayarit (2018)

In October 2018, Hurricane Willa hit 8 of the 20 municipalities in Nayarit State. The climate phenomenon caused the flooding of the San Pedro and Acaponeta rivers, which affected more than 180,000 people [55]. Willa struck the municipality of San Blas on October 24, but their effects were felt a few days before. For this event, the dates envisaged for the periods are as follows: from 10-14-2018 to 10-19-2018 for before, from 10-20-2018 to 10-26-2018 for **during**, and from 11-10-2018 to 11-16-2018 for **after**. Fig. 7 depicts some images representative of the event. 50 images were collected for each time period for this event.



Fig. 6. Protests during International Women's Day in Mexico City.



Fig. 7. Impacts of Hurricane Willa in San Blas, Nayarit.

4. Protesters in Guadalajara During 2020

On 4 May 2020, a 30-year-old young man, Giovanni López was arrested for assaulting a group of police officers from the municipality of Ixtlahuacán, Jalisco. The next day, he died of damage caused by the police [56]. These events led to civil mobilizations in Guadalajara on days 4, 5, and 6 of June. The dates considered for the periods are: from 05-28-2020 to 05-30-2020 for before, from 06-04-2020 to 06-06-2020 for during, and from 07-01-2020 to 07-03-2020 for after. Fig. 8 presents some representative images. 100 images were collected for each time period for this event.



Fig. 8. Protests in June 2020 at Guadalajara, Jalisco.

5. Violence in Acapulco During 2017

In 2017, the municipality of Acapulco was considered the most violent place for women according to the United Nations. According to the report, the municipality contributed 3.9% of the national homicides, 5.5% above the national average [57]. This wave of violence is still far from over, considering the health contingency of 2020, we think that with fewer people on the streets and the beaches closed, a change can be identified by comparing similar time periods. Therefore, we compared images from the months of April, May, and June in the years 2017 and 2020. We chose these months because of increased tourist arrivals during the warmer months. The representative pictures are in Fig. 9. 50 images were collected for each time period for this event.



Fig. 9. Images of Acapulco during 2017 and 2020.

B. Reference Pictures

To be able to identify the differences, we also need a reference base for the tourist destinations we want to analyze. To this end, we have collected around 500 photographs of the destinations where the events took place. As reference material, we have used photographs from local DMOs (SECTUR) to promote the destination, and to ensure variety we also used material from Flickr. The selection was done manually to ensure that the photos were from the place of interest and to avoid duplicated or bad-quality material. For our approach, there are no limits to the number of photographs to analyze. The reason behind the choice of 500 photographs is strictly linked to the variety and quality of the material at hand. Fig. 10 shows a set of representative pictures of each considered destination.



Fig. 10. Sample of the reference images of the destinations under analysis.

V. EXPERIMENTAL RESULTS

Three different approaches were used for the ground distance required to measure the similarities between histograms. As mentioned

above, two distance matrices were obtained through NLP approaches (WordNet and Embeddings), as a substitute for language experts. The final method was to consider the difference between all the concepts as 1 unit. Fig. 11, Fig. 12, Fig. 13, Fig. 14 and Fig. 15 show the graphs of the calculated difference between the histogram of the reference images and the histograms of the set of pictures on each time frame. There is a chart for each approach adopted for the distance matrix (WordNet, Embeddings, and Distance = 1).

From the figures, we can highlight the existence of a triangular pattern in Cancun, Mexico City, San Blas, and Guadalajara. The *peak* vertex belongs to the “during” period in all these charts, and is pointing up in San Blas, and downwards in Cancun, Mexico City, and Guadalajara. The existence of this triangular pattern confirms our hypothesis that during abnormal conditions such as adverse events, the differences between the baseline images (those from the DMOs) and the ones from uncontrolled sources are incremented. On the other hand, despite there being differences between the baseline photos and those from normal conditions, they are shorter than the above. In this context, we refer to normal conditions as the time intervals previous to and posterior to an adverse event for the destination image. The intuition behind this idea is that photographs for such “normal conditions” present the destination with all its pros and cons while during adverse events, photographs reflect more flaws than qualities. Given that controlled sources present just the brighter side of the destination they promote, photographs of non-anomalous periods (posted by uncontrolled sources) will also have a variation regarding the image projected by DMOs but smaller than those resulting from adverse events.

As for Acapulco’s charts, it should be noted that the series for 2017 and 2020 shows differences in magnitudes and are separable. It is important to note that the patterns are similar to the three ways to compute the distance matrix. The magnitudes obtained with WordNet and Embeddings are almost the same, and with distance = 1, the differences between the stages are more evident and, therefore the one that we will discuss in this section.

Even if the triangular model in the series is not pointing in the same direction in all destinations, the anomaly is easily identifiable. Table II and Table III present the score given by EMD to each period, the average score of a series, and the rate of change in the critical point given the mean. The rate of change (last column of both tables) presents a statistic of how much varies the score from periods considered “normal” and anomalous periods. This measure can help us understand the level of impact of factors depicted by photos during adverse events. Table II shows that the most drastic change occurred in San Blas (from the destinations in this group), with a change rate of 10.731%. Considering the damages inflicted by Hurricane Willa on this town, it is understandable those variations. Regarding Mexico City and Guadalajara, the damage provoked by the protesters is well reflected in the media and has resulted in significant changes in their scores, with a change rate of 8.876% and 8.299% each. Cancun got the smallest change rate in this comparison (4.115%). Given that Sargassum is just a small flaw in the landscape, this rate of change makes sense. Data from Acapulco (Table III) shows an important separation in the periods evaluated. This situation could be due to pictures in 2017 presenting noise and festive scenes, with different examples of crime and law enforcement. On the other hand, in 2020, with quarantine and closed beaches, images of solitude in the landscape are free of trash and jolly chaos.

It is also interesting to know what attributes were found by the Neural Networks in the scenes corresponding to the critical points. By using this information and the changes in the series, our approach can provide an overview of the current situation. Table IV shows some scene attributes that stand out in critical points, and Table V

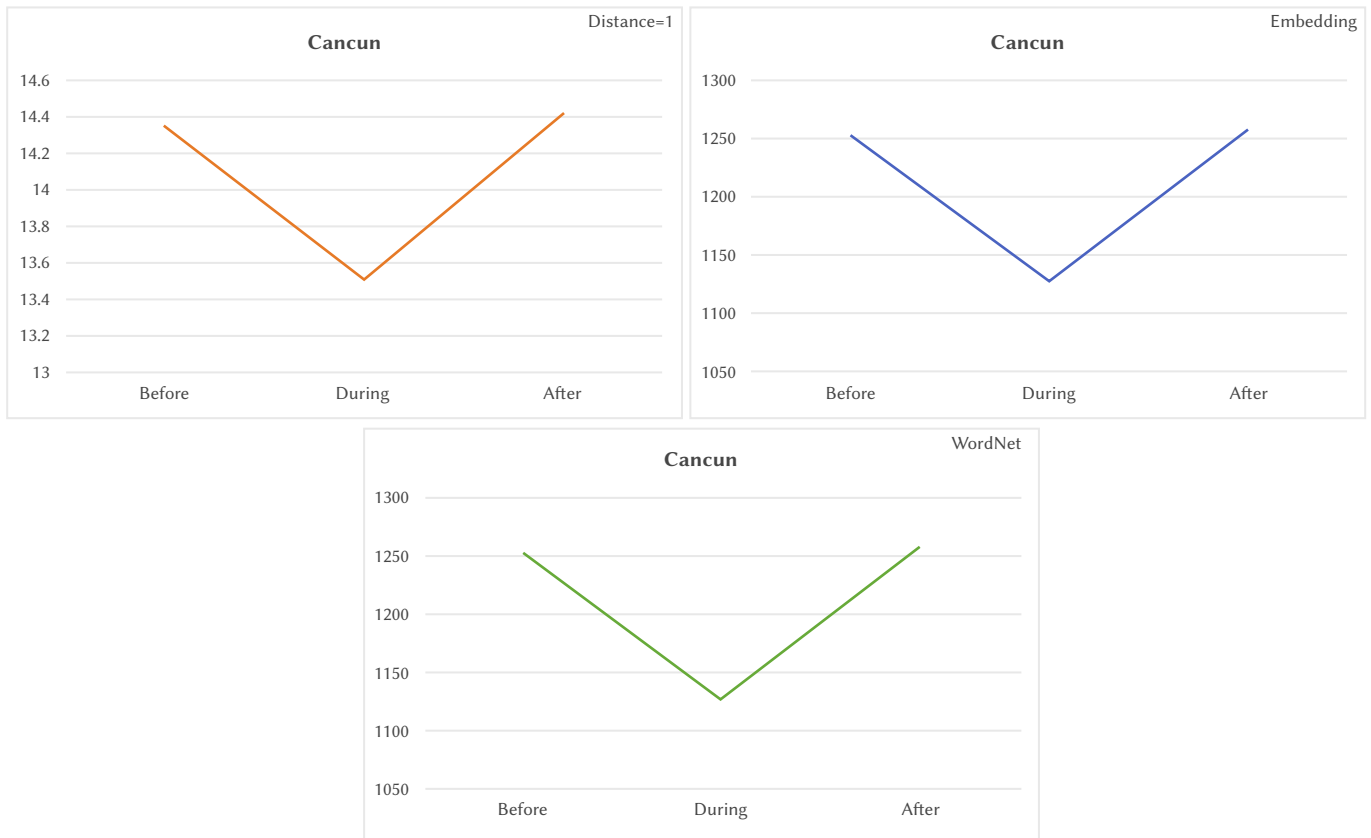


Fig. 11. Charts for Cancun. Differences between histograms at each time frame. In the upper right corner, a label is given for each method.



Fig. 12. Charts for Mexico City. Differences between histograms at each time frame. In the upper right corner, a label is given for each method.



Fig. 13. Charts for San Blas. Differences between histograms at each time frame. In the upper right corner, a label is given for each method.



Fig. 14. Charts for Guadalajara. Differences between histograms at each time frame. In the upper right corner, a label is given for each method.

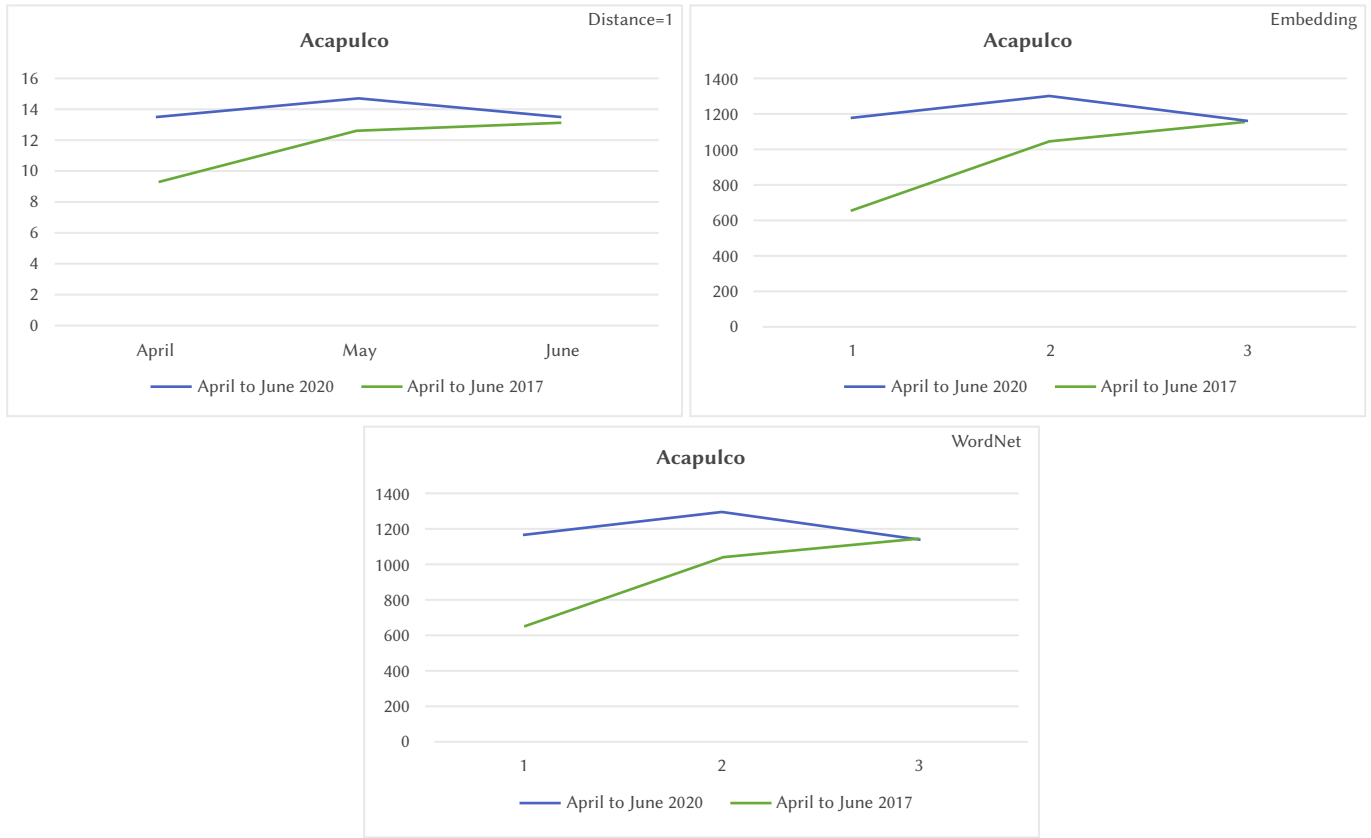


Fig. 15. Charts for Acapulco. Differences between histograms at each time frame. In the upper right corner, a label is given for each method.

TABLE II. SCORES FOR THE DIFFERENCES BETWEEN HISTOGRAMS (REFERENCE AND PERIOD) WITH GROUND-DISTANCE = 1 FOR ALL THE ATTRIBUTES OF THE SCENE

Destination/Period	Before	During	After	Average	Rates of change in the <i>During</i> period
Cancun	14.348	13.513	14.418	14.093	4.115%
Mexico City	12.542	10.905	12.455	11.967	8.876%
San Blas	9.102	10.583	8.987	9.557	10.731%
Guadalajara	18.538	16.26	18.397	17.731	8.299%

TABLE III. SCORES FOR THE DIFFERENCE BETWEEN THE HISTOGRAMS OF ACAPULCO (REFERENCE AND PERIOD). THE GROUND-DISTANCE = 1

Month/Year	2017	2020	Difference	Rates of change in the months considered
April	9.359	13.569	4.210	36.723%
May	12.688	14.764	2.076	15.124%
June	13.240	13.574	0.334	2.491%

presents those attributes that were not present in the scenes of the *during* period but are part of the scenes of the reference photographs for all the destinations. The attributes were obtained through set operations between the items in the reference set and the items in the period under evaluation. It can be seen that the most common faults found by neural networks are *dirt* and *dirty*. These attributes are understandable for mass manifestations, sargassum, crime, and climatological phenomena, however, defects in scenes for Hurricane Willa have been detected by neural networks as bathing scenes. As for the negative image (obtained with the missing attributes in Table V), it can be observed that such elements can tell us about things that are absent in the photographs of a destination. For example, for Cancun, all the photos were aimed at showing the algae problem and did not reflect the fun things about the beaches as surfing or soothing. The same applies to San Blas with pictures showing the damage caused by the hurricane. For better understanding, Table VI shows the complete list of attributes identified in the scenes of the reference images in all

the destinations considered. The Acapulco case is interesting, as the chaotic images of people enjoying the beach and those photos that portray law enforcement present different realities side by side, but also show all of Acapulco's strengths at the same time. This series presented a violent season for our country and got a better score by EMD instead of the one that presents the loneliness of the quarantine (series 2020) because the last one presented only a few aspects of the place. We believe that this is one of the reasons behind the best score in the 2017 series. The other and most important reason is that, like in many scene recognition tools, the deep learning module was trained to recognize a wide range of elements, but none is unpleasing to the eye. That's why problems like crime go unnoticed by neural networks, and flood scenes get confused with people bathing. However, despite these flaws, our method can assess the disparity between photographs to promote a destination and those of uncontrolled sources, giving details of attributes that are not reflected in a critical point or the defects that are present in the period under assessment.

TABLE IV. SCENE ATTRIBUTES THAT STAND OUT IN THE DURING PERIOD IN ALL DESTINATIONS

Destination	Attributes
Cancun	Dirty, dirt.
Mexico City	Dirty, dirt, cluttered space, waiting in line, stressful.
San Blas	Bathing, cold, pavement.
Guadalajara	Dirty, dirt, spectating.
Acapulco (2017)	Dirty, dirt, stressful, shingles.
Acapulco (2020)	Dirty, dirt, rusty.

TABLE V. ATTRIBUTES THAT WERE NOT PRESENT IN THE SCENES OF THE DURING PERIOD FOR THE DIFFERENT DESTINATIONS

Destination	Attributes
Cancun	Surf, competing, shopping, soothing, vegetation. Total: 29 attributes missing.
Mexico City	Symmetrical, soothing, playing, matte. Total: 18 attributes missing.
San Blas	Brick, surf, aged, soothing, ocean. Total: 17 attributes missing.
Guadalajara	Brick, soothing, glossy, concrete, symmetrical. Total: 33 attributes missing.
Acapulco (2017)	Surf, symmetrical, soothing, camping, glossy, sports, hiking, horizontal components. Total: 41 attributes missing in the entire series.
Acapulco (2020)	Symmetrical, glossy, rock, playing, grass, shopping, climbing, camping. Total: 64 attributes missing in the entire series.

VI. DISCUSSION

According to scores in each time frame for the three employed distances on the destinations, there was a significant difference for San Blas, where the unfortunate event of the Willa Hurricane significantly increased images from uncontrolled sources. In this regard, the DIT of

San Blas continues forming through positive and negative images [10], [11]. Also the DIT is related to a set of ideas, impressions, and beliefs about a particular place (San Blas in this case) based on medium to long-term experiences [9], where images from uncontrolled sources play a major role. Thus, the increment of pictures from adverse events impacts the "induced" level (one of the three levels of destination image formation analysis) [12]. Since such images are intentionally published by different uncontrolled media, they influence the processing of the destination image, even if it is based on negative aspects of the destination promotion. Consequently, once the adverse event has occurred, DMOs should resume promotion with positive images intended for potential visitors, particularly cognitive resources that impact more [16]. The aforementioned is also related to perceived changes during adverse economic events where the construction of such destination image changes the decisions of potential visitors [23]–[25].

Concerning the measures given by our approach, differences among histograms show that the rate of change in each critical point, considering the mean between normal and abnormal periods, contributes to understanding the level of influence of the factors behind the analyzed pictures of adverse events. Of the destinations analyzed, San Blas represented the higher rate of change of the places in this study for its images (Huracan Willa), followed by Mexico City (protestors, vandalism), Guadalajara (protestors, vandalism), and Cancun (sargassum). It can be said that adverse periods affecting the DIT by exogenous factors hurt this image [40], which presented a method that measures the dissimilarity between images of different destinations influenced by endogenous and exogenous factors and helps them identify the elements that influence positively or negatively the destination image. In this case, San Blas was severely damaged by the hurricane, and our approach was able to detect the change based on the comparative analysis of the photographs before, during, and after the event. The scene recognition approach that takes into account objects, environment, and context, is therefore preferred over simple object detection frameworks [43].

TABLE VI. ATTRIBUTES OF THE SCENE FROM THE SET OF REFERENCE PHOTOS

Destination	Attributes
Cancun	Surf, dry, horizontal components, cold, biking, cluttered space, glass, enclosed area, cloth, praying, plastic, socializing, sand, rugged scene, eating, railing, competing, shopping, glossy, semi-enclosed area, soothing, natural light, open area, pavement, spectating, vertical components, man-made, wood, indoor lighting, touring, moist, metal, warm, driving, transporting, shrubbery, congregating, vegetation, sports, grass, natural, asphalt, boating, ocean, diving, swimming, far-away horizon, stillwater, sunny, foliage, leaves, clouds, trees, reading, no horizon, aged, climbing.
Mexico City	Warm, fencing, rusty, man-made, cold, semi-enclosed area, sports, glossy, congregating, rugged scene, matte, symmetrical, carpet, dirty, paper, dirt, horizontal components, metal, working, railing, plastic, transporting, soothing, rock, shingles, ocean, playing, stressful, flowers, wire, shopping, driving, swimming, boating, glass, brick, natural light, asphalt, biking, pavement, wood, aged, far-away horizon, cluttered space, socializing, dry, still water, diving, cloth, eating, competing, natural, climbing, shrubbery, touring, praying, using tools, enclosed area, clouds, open area, spectating, moist, no horizon, indoor lighting, leaves, trees, camping, vertical components, vegetation, sunny, foliage.
San Blas	Glass, brick, using tools, indoor lighting, working, dirt, rock, driving, vertical components, transporting, praying, asphalt, wood, natural light, dry, warm, metal, dirty, climbing, surf, cluttered space, open area, shrubbery, grass, shingles, man-made, soothing, shopping, enclosed area, touring, aged, sunny, camping, moist, far-away horizon, congregating, boating, swimming, clouds, no horizon, natural, diving, leaves, cloth, rugged scene, foliage, still water, semi-enclosed area, vegetation, ocean, trees, biking.
Guadalajara	Diving, rusty, metal, paper, brick, soothing, man-made, competing, still water, eating, glossy, congregating, plastic, wood, semi-enclosed area, glass, using tools, scary, railing, concrete, boating, stressful, natural light, socializing, working, waiting in line, moist, dirty, dirt, aged, symmetrical, shopping, camping, cloth, asphalt, horizontal components, carpet, shingles, transporting, driving, dry, warm, pavement, indoor lighting, natural, swimming, open area, climbing, fencing, shrubbery, biking, sand, praying, trees, cluttered space, leaves, rock, touring, grass, enclosed area, vegetation, hiking, far-away horizon, foliage, reading, rugged scene, vertical components, clouds, no horizon, sunny, sports, running water.
Acapulco	Reading, congregating, semi-enclosed area, pavement, cluttered space, socializing, competing, dirty, glass, symmetrical, aged, sand, metal, glossy, natural light, open area, cold, driving, plastic, eating, biking, sports, shopping, horizontal components, playing, asphalt, working, paper, railing, hiking, cloth, enclosed area, praying, warm, wood, using tools, grass, bathing, touring, climbing, transporting, vertical components, spectating, surf, indoor lighting, dry, soothing, foliage, rugged scene, camping, leaves, still water, man-made, rock, vegetation, natural, diving, boating, moist, trees, ocean, shrubbery, dirt, swimming, far-away horizon, sunny, no horizon, clouds.

In the Acapulco analysis, the highest rates of change were observed in April, followed by May and June for 2017 and 2020. Considering that in March of 2017, the Tourism tianguis¹ took place in such destination, promotion images presented the beautiful sides of the city such as natural landscapes, urban landscapes, and related tourist products; Compared to the events that took place in the following months of the same year (insecurity, crime, social agglomerations) and the events of 2020 at the beginning of the SARS CoV-2 pandemic (closed beaches, minimal hotel occupancy, emergency health measures), produced a different picture of the destination. Once more, the scene recognition approach made it possible to provide descriptions of the image and its content that have been useful in distinguishing the observed features presented on images from controlled and uncontrolled sources.

As regards the attributes of critical points in all cities, two coincidences have been noted: “dirty” and “dirt”; Representative characteristics of the identified adverse events (protestors, vandalism, delinquency, sargassum, and climatic factors). Regarding San Blas and the impact of the hurricane, the prominent attributes were “bathing”, “cold”, and “pavement”. Such attributes were misinterpreted by artificial intelligence but they pointed out inundations in the scene, which are negative factors for the destination promotion but that are considered by potential tourists. In this sense, our methodology could improve attribute recognition (or minimize the mistakes) by nesting other specialized computer vision strategies, e.g. [58] [59].

The choice of a destination is strongly influenced by different secondary factors [13], therefore, being able to measure an important part of them, communicated through powerful and simple cognitive resources like the photographs, could help DMOs and other entities to plan countermeasures to mitigate the effects of adverse events for the tourism industry.

VII. CONCLUSIONS

In this work, we presented an approach based on deep learning techniques to automatically measure changes between the photographs used to promote a destination by the DMOs (controlled sources) and those published on the Internet (uncontrolled sources) during adverse events that negatively could influence the projected destination image. This approach was designed to automate the tracking of these changes, reducing the burden on experts or to be employed in the absence of them.

Our method uses different computing techniques and employs the Deep Learning paradigm as a cornerstone. Thanks to the synergy of these techniques, our method can detect critical points and also provide information about the scene attributes that stand out in such events. With this information, travelers and/or DMOs can design strategies to attenuate the factors that may discourage potential tourists from visiting a destination. To our knowledge, this is the first work to explore this issue through the use of IT techniques.

As our proposal is a pioneer in dealing with the problem of measuring differences between images from controlled sources versus images from uncontrolled sources, specifically during events that could alter the projected destination image, it is possible that our methodology could be easily adapted to other potential applications, for example:

- Destination development (Dd): Given a set of parameters, the present methodology can be adapted to measure the evolution of a destination’s infrastructure using a collection of images. This can also be used for the sustainable development of the destination.

- Recommendation systems (RS): Given the change history tracking in a period over some destination, it is possible to use this information to feedback into an RS to support tourists to avoid unpleasant experiences and to recommend the best periods to travel, minimizing witnessing undesirable situations.
- Online Travel Reviews (OTRs): The analysis of the OTRs has been studied a lot, mainly using Sentiment Analysis. However, the OTRs analysis does not differentiate between normal experience periods and those with adverse events. Thus our approach can be used to segment OTRs depending on the detected changes and have a deeper OTRs analysis.
- Quick response (QS): Taking into account historical records, DMOs, and governments can discover patterns and design strategies to mitigate drawbacks in advance.
- Ranking Systems (RaS): Destinations can be ranked according to their variations in the monitoring of changes. Destinations with fewer variations (according to a baseline) are ranked above those with several variations.
- Measurement of differences in content (MDC): Our methodology can be used to quantify differences between UGC and DMOs’ content using photos. With this measure, DMOs can evaluate the performance of their marketing campaigns regarding the experiences shared by tourists.

Finally, the proposed method provided evidence that images can be used to measure differences in the destination image projected by controlled and uncontrolled sources, it can be used for different destinations and it helps to detect critical points on the tracking. Using this method as a cornerstone, our investigation will be able to go to further steps to analyze the impact that those divergences have on tourists’ destination image through in situ analyses.

REFERENCES

- [1] UNWTO, *Panorama del turismo internacional, Edición 2020*. World Tourism Organization (UNWTO), 2021.
- [2] A. Feizollah, M. M. Mostafa, A. Sulaiman, Z. Zakaria, A. Firdaus, “Exploring halal tourism tweets on social media,” *Journal of Big Data*, vol. 8, no. 1, pp. 1–18, 2021.
- [3] B. Toumeh, “Exploitation of digital transformation technologies in smart tourism destinations: Facts and challenges,” *Digital Transformation: IoT, AI, VR, Big Data*, p. 124, 2021.
- [4] B. T. Khoa, N. M. Ly, V. T. T. Uyen, N. T. T. Oanh, B. T. Long, “The impact of social media marketing on the travel intention of z travelers,” in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2021, pp. 1–6, IEEE.
- [5] M. T. Cuomo, D. Tortora, P. Foroudi, A. Giordano, G. Festa, G. Metallo, “Digital transformation and tourist experience co-design: Big social data for planning cultural tourism,” *Technological Forecasting and Social Change*, vol. 162, p. 120345, 2021, doi: 10.1016/j.techfore.2020.120345.
- [6] F. A. C. Calderón, M. V. V. Blanco, “Impacto de internet en el sector turístico,” *Revista UNIANDES Episteme*, vol. 4, no. 4, pp. 477–490, 2017.
- [7] L. Lalicic, A. Huertas, A. Moreno, M. Jabreel, “Emotional brand communication on facebook and twitter: Are dmos successful?,” *Journal of Destination Marketing & Management*, vol. 16, p. 100350, 2020, doi: 10.1016/j.jdmm.2019.03.004.
- [8] M. del Mar Gálvez-Rodríguez, J. Alonso-Cañadas, A. H. de Rosario, C. Caba-Pérez, “Exploring best practices for online engagement via facebook with local destination management organisations (dmos) in europe: A longitudinal analysis,” *Tourism Management Perspectives*, vol. 34, p. 100636, 2020, doi: 10.1016/j.tmp.2020.100636.
- [9] J. L. Crompton, “Motivations for pleasure vacation,” *Annals of Tourism Research*, vol. 6, no. 4, pp. 408–424, 1979, doi: 10.1016/0160-7383(79)90004-5.
- [10] S. Choi, X. Y. Lehto, A. M. Morrison, “Destination image representation on the web: Content analysis of macau travel related websites,” *Tourism management*, vol. 28, no. 1, pp. 118–129, 2007.

¹ Is an international event for tourism promotion of Mexican destinations held by the federal government.

- [11] J. Li, F. Ali, W. Kim, "Reexamination of the role of destination image in tourism: An updated literature review," *e-Review of Tourism Research*, vol. 12, pp. 191–209, 2015.
- [12] C. Gunn, *Vacationscape: Developing Tourist Areas*. Taylor & Francis, 1997.
- [13] M. R. González-Rodríguez, R. Martínez-Torres, S. Toral, "Post-visit and pre-visit tourist destination image through eWOM sentiment analysis and perceived helpfulness," *International Journal of Contemporary Hospitality Management*, vol. 28, no. 11, pp. 2609–2627, 2016, doi: 10.1108/ijchm-02-2015-0057.
- [14] S. Baloglu, K. W. McCleary, "A model of destination image formation," *Annals of Tourism Research*, vol. 26, no. 4, pp. 868–897, 1999, doi: 10.1016/S0160-7383(99)00030-4.
- [15] A. Beerli, J. D. Martín, "Factors influencing destination image," *Annals of Tourism Research*, vol. 31, no. 3, pp. 657–681, 2004, doi: 10.1016/j.annals.2004.01.010.
- [16] D. M. Frías, M. A. Rodríguez, J. A. Castañeda, "Internet vs. travel agencies on pre-visit destination image formation: An information processing view," *Tourism Management*, vol. 29, no. 1, pp. 163–179, 2008, doi: 10.1016/j.tourman.2007.02.020.
- [17] M. G. Gallarza, I. G. Saura, H. C. García, "Destination image," *Annals of Tourism Research*, vol. 29, no. 1, pp. 56–78, 2002, doi: 10.1016/S0160-7383(01)00031-7.
- [18] R. Govers, F. M. Go, "Deconstructing destination image in the information age," *Information Technology & Tourism*, vol. 6, no. 1, pp. 13–29, 2003, doi: 10.3727/109830503108751199.
- [19] B. Bramwell, L. Rawding, "Tourism marketing images of industrial cities," *Annals of Tourism Research*, vol. 23, no. 1, pp. 201–221, 1996, doi: 10.1016/0160-7383(95)00061-5.
- [20] W. C. Gartner, "Image formation process," *Journal of Travel & Tourism Marketing*, vol. 2, no. 2-3, pp. 191–216, 1994, doi: 10.1300/j073v02n02_12.
- [21] J.-R. Chang, M.-Y. Chen, L.-S. Chen, S.-C. Tseng, "Why customers don't revisit in tourism and hospitality industry?," *IEEE Access*, vol. 7, pp. 146588–146606, 2019, doi: 10.1109/access.2019.2946168.
- [22] M. Nowacki, A. Niezgoda, "Identifying unique features of the image of selected cities based on reviews by TripAdvisor portal users," *Scandinavian Journal of Hospitality and Tourism*, vol. 20, no. 5, pp. 503–519, 2020, doi: 10.1080/15022250.2020.1833362.
- [23] M. qi Cao, J. Liang, M. zhao Li, Z. hao Zhou, M. Zhu, "TDIVis: visual analysis of tourism destination images," *Frontiers of Information Technology & Electronic Engineering*, vol. 21, no. 4, pp. 536–557, 2020, doi: 10.1631/fitee.1900631.
- [24] S. Stepchenkova, A. P. Kirilenko, E. Shichkova, "Influential factors for intention to visit an adversarial nation: increasing robustness and validity of findings," *International Journal of Tourism Cities*, vol. 5, no. 3, pp. 491–510, 2019, doi: 10.1108/IJTC-11-2018-0085.
- [25] E. Marine-Roig, B. Ferrer-Rosell, N. Daries, E. Cristobal-Fransi, "Measuring gastronomic image online," *International Journal of Environmental Research and Public Health*, vol. 16, no. 23, p. 4631, 2019, doi: 10.3390/ijerph16234631.
- [26] S. Song, K. Park, et al., "Thematic analysis of destination images for social media engagement marketing," *Industrial Management & Data Systems*, vol. ahead- of-print, no. ahead-of-print, 2020, doi: 10.1108/IMDS-12-2019-0667.
- [27] M. T. Liu, Y. Liu, Z. Mo, K. L. Ng, "Using text mining to track changes in travel destination image: the case of macau," *Asia Pacific Journal of Marketing and Logistics*, vol. 33, no. 2, pp. 371–393, 2020, doi: 10.1108/APJML-08-2019-0477.
- [28] R. Wang, J. Luo, S. S. Huang, "Developing an artificial intelligence framework for online destination image photos identification," *Journal of Destination Marketing & Management*, vol. 18, p. 100512, 2020, doi: 10.1016/j.jdmm.2020.100512.
- [29] E. Marine-Roig, A. Huertas, "How safety affects destination image projected through online travel reviews," *Journal of Destination Marketing & Management*, vol. 18, p. 100469, 2020, doi: 10.1016/j.jdmm.2020.100469.
- [30] L. B. Ferreira, J. d. M. E. Giraldi, "Rio de janeiro's image as the 2016 olympic games host city: analysis of the main image formation factors," *Journal of Hospitality and Tourism Insights*, vol. 3, no. 2, pp. 115–135, 2020, doi: 10.1108/JHTI-03-2019-0037.
- [31] K. Zhang, Y. Chen, Z. Lin, "Mapping destination images and behavioral patterns from user-generated photos: a computer vision approach," *Asia Pacific Journal of Tourism Research*, vol. 25, no. 11, pp. 1199–1214, 2020, doi: 10.1080/10941665.2020.1838586.
- [32] L. J. Nixon, "An online image annotation service for destination image measurement," *e-Review of Tourism Research*, vol. 17, no. 2, 2019.
- [33] G. Sun, "Symmetry analysis in analyzing cognitive and emotional attitudes for tourism consumers by applying artificial intelligence python technology," *Symmetry*, vol. 12, no. 4, p. 606, 2020, doi: 10.3390/sym12040606.
- [34] S. L. Toral, M. R. Martínez-Torres, M. R. Gonzalez- Rodriguez, "Identification of the unique attributes of tourist destinations from online reviews," *Journal of Travel Research*, vol. 57, no. 7, pp. 908–919, 2017, doi: 10.1177/0047287517724918.
- [35] R. Micera, R. Crispino, "Destination web reputation as "smart tool" for image building: the case analysis of naples city-destination," *International Journal of Tourism Cities*, vol. 3, no. 4, pp. 406–423, 2017, doi: 10.1108/ijtc-11-2016-0048.
- [36] C. H. Chin, M. C. Lo, Z. bin Razak, P. Pasbakhsh, A. A. Mohamad, "Resources confirmation for tourism destinations marketing efforts using PLS-MGA: The moderating impact of semirural and rural tourism destination," *Sustainability*, vol. 12, no. 17, p. 6787, 2020, doi: 10.3390/su12176787.
- [37] R. Leung, H. Q. Vu, J. Rong, "Understanding tourists' photo sharing and visit pattern at non-first tier attractions via geotagged photos," *Information Technology & Tourism*, vol. 17, no. 1, pp. 55–74, 2017, doi: 10.1007/s40558-017-0078-3.
- [38] L. J. B. Nixon, "An online image annotation service for destination image measurement," *e-Review of Tourism Research*, vol. 17, no. 2, 2019.
- [39] X. Xiao, C. Fang, H. Lin, "Characterizing tourism destination image using photos' visual content," *ISPRS International Journal of Geo-Information*, vol. 9, no. 12, p. 730, 2020, doi: 10.3390/ijgi9120730.
- [40] A. Arabadzhyan, P. Figini, L. Vici, "Measuring destination image: a novel approach based on visual data mining. a methodological proposal and an application to european islands," *Journal of Destination Marketing & Management*, vol. 20, p. 100611, 2021, doi: 10.1016/j.jdmm.2021.100611.
- [41] V. Bui, A. R. Alaei, H. Q. Vu, G. Li, R. Law, "Revisiting tourism destination image: A holistic measurement framework using big data," p. 004728752110247, 2021, doi: 10.1177/00472875211024749.
- [42] Z. He, N. Deng, X. R. Li, H. Gu, "How to "read" a destination from images? machine learning and network methods for DMOs' image projection and photo evaluation," p. 004728752199513, 2021, doi: 10.1177/0047287521995134.
- [43] L. Xie, F. Lee, L. Liu, K. Kotani, Q. Chen, "Scene recognition: A comprehensive survey," *Pattern Recognition*, vol. 102, p. 107205, 2020, doi: 10.1016/j.patcog.2020.107205.
- [44] A. Oliva, A. Torralba, "The role of context in object recognition," *Trends in Cognitive Sciences*, vol. 11, no. 12, pp. 520–527, 2007, doi: 10.1016/j.tics.2007.09.009.
- [45] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018, doi: 10.1109/tpami.2017.2723009.
- [46] Y. Tang, L. H. U, Y. Cai, N. Mamoulis, R. Cheng, "Earth mover's distance based similarity search at scale," *Proceedings of the VLDB Endowment*, vol. 7, no. 4, pp. 313–324, 2013, doi: 10.14778/2732240.2732249.
- [47] G. A. Miller, *WordNet: An electronic lexical database*. MIT press, 1998.
- [48] Z. Wu, M. Palmer, "Verb Semantics and Lexical Selection," *arXiv e-prints*, pp. cmp-1g/9406033, 1994.
- [49] T. Pedersen, S. Patwardhan, J. Michelizzi, et al., "Wordnet:: Similarity-measuring the relatedness of concepts," in *AAAI*, vol. 4, 2004, pp. 25–29.
- [50] T. Mikolov, W.-t. Yih, G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, 2013, pp. 746–751, Association for Computational Linguistics.
- [51] T. Mikolov, "Google code archive - long- term storage for google code project hosting," 2013. [Online]. Available: <https://code.google.com/archive/p/word2vec/>.
- [52] SECTUR, "Compendio estadístico del turismo en México 2020," 2021. [Online]. Available: <https://datos.sectur.gob.mx>.
- [53] D. Rendón, "Sargazo en Cancún: Por qué llega y en qué temporada hay más," 2019. [Online]. Available: <https://tipsparatuviaje.com/sargazo-en-cancun/>.

- [54] BBC, "Women's day: Protesters clash with police in Mexico," 2021. [Online]. Available: <https://www.bbc.com/news/world-latin-america-56329666>.
- [55] M. A. Navarro-Quintero, *Huracan Willa y sus efectos en Nayarit*, vol. 1 of 1. Senado de La Republica, 1 ed., 2018.
- [56] J. Martínez, "Lo que sabemos de caso Giovanni López, detenido en Ixtlahuacán," 2020. [Online]. Available: <https://www.milenio.com/policia/giovanni-lopez-asesinado-jalisco-mexico-cubrebocas>.
- [57] CNN, "Acapulco, la ciudad más peligrosa de México para las mujeres, según la ONU," 2017. [Online]. Available: <https://cnnespanol.cnn.com/2017/12/28/acapulco-la-ciudad-mas-peligrosa-de-mexico-para-las-mujeres-segun-la-onu/>.
- [58] W. Han, L. Cao, S. Xu, "A method of the coverage ratio of street trees based on deep learning," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 5, p. 23, 2022, doi: 10.9781/ijimai.2022.07.003.
- [59] Z. H. Arif, M. Mahmoud, K. H. Abdulkareem, S. Kadry, M. A. Mohammed, M. N. Al-Mhiqani, A. S. Al-Waisy, J. Nedoma, "Adaptive deep learning detection model for multi-foggy images," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 7, p. 26, 2022, doi: 10.9781/ijimai.2022.11.008.



Angel Díaz-Pacheco

Angel Díaz-Pacheco received his degree in Computer Systems Engineering from the Technological Institute of Veracruz in 2010, his Master's degree in Computer Systems from the Technological Institute of Apizaco in 2013, and his Doctorate in Science in the area of Computer Science from INAOE in 2019. He completed a postdoctoral stay at the Technology Transfer Unit of CICESE-UT3. He is currently a full-time professor at the Faculty of Engineering at the University of Guanajuato. He belongs to the National System of Researchers level I. His research interests include machine learning, big data, soft computing, fuzzy logic, and their application in the study of significant constructs in tourism research.



Miguel A. Álvarez-Carmona

Miguel A. Álvarez-Carmona obtained his master's and doctorate degrees in Computational Sciences in 2014 and 2019, respectively, from the National Institute of Astrophysics, Optics, and Electronics in Mexico. Currently, he is a researcher at CIMAT, where his main line of research is the application of artificial intelligence to tourism. He also belongs to the National System of Researchers, he also is a member of the Mexican Association of Natural Language Processing and the Mexican Association of Artificial Intelligence.



Ansel Y. Rodríguez-González

Ansel Y. Rodríguez-González earned his Bachelor's degree in Computer Science and his Master's degree in Mathematics from Havana University, Cuba, in 2004 and 2007, respectively; and his Ph.D. in Computer Science at the National Institute of Astrophysics, Optics, and Electronics in México. He is a CONAHcyT-researcher at CICESE-UT3 and he holds the distinction of being a Level 1 member of the National System of Researchers. His current research interests include data mining, pattern recognition, machine learning, and evolutionary computation.



Hugo Carlos

Hugo Carlos is a researcher at the Center for Geospatial Information Sciences A.C. He earned both his Master's and Ph.D. in Sciences from the Center for Research in Mathematics. Currently, he is an active member of the National Laboratory for Geointelligence and also holds the distinction of being a Level 1 Member of the National System of Researchers. His research focus spans a range of specialized topics, including machine learning and deep learning, remote sensing algorithms, image processing, and optimization.



Ramón Aranda

Ramón Aranda earned his bachelor's degree in Computer Science from the University of Yucatán and received a master's degree and a Ph.D. in Computer Science from the Centro de Investigación en Matemáticas (CIMAT), since 2010 and 2016 respectively. Currently, he is a CONAHcyT-Researcher at CIMAT, Mexico. He belongs to the Mexican National System of Researchers (Level I) and is member of the Tourism Research Mexican Academy. His research interests are Tourism Data Analysis, Computer Vision, Optimization, Machine Learning and Data Science.

An Adaptive Salp-Stochastic-Gradient-Descent-Based Convolutional LSTM With MapReduce Framework for the Prediction of Rainfall

S. Oswalt Manoj¹, Abhishek Kumar², Ashutosh Kumar Dubey^{3*}, J. P Ananth¹

¹ Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Kuniyamuthur, Coimbatore, Tamil Nadu (India)

² Department of Computer Science and Engineering, Chandigarh University, Gharuan, Mohali, Punjab (India)

³ Chitkara University School of Engineering and Technology, Chitkara University, Himachal Pradesh (India)

* Corresponding author: ashutosh.dubey@chitkara.edu.in

Received 4 September 2022 | Accepted 22 September 2023 | Early Access 22 January 2024



ABSTRACT

Rainfall prediction is considered to be an esteemed research area that impacts the day-to-day life of Indians. The predominant income source of most of the Indian population is agriculture. It helps the farmers to make the appropriate decisions pertaining to cultivation and irrigation. The primary objective of this investigation is to develop a technique for rainfall prediction utilising the MapReduce framework and the convolutional long short-term memory (ConvLSTM) method to circumvent the limitations of higher computational requirements and the inability to process a large number of data points. In this work, an adaptive salp-stochastic-gradient-descent-based ConvLSTM (adaptive S-SGD-based ConvLSTM) system has been developed to predict rainfall accurately to process the long time series data and to eliminate the vanishing problems. To optimize the hyperparameter of the convLSTM model, the S-SGD methodology proposed combine the SGD and the salp swarm algorithm (SSA). The adaptive S-SGD based ConvLSTM has been developed by integrating the adaptive concept in S-SGD. It tunes the weights of ConvLSTM optimally to achieve better prediction accuracy. Assessment measures, such as the percentage root mean square difference (PRD) and mean square error (MSE), were employed to compare the suggested method with previous approaches. The developed system demonstrates high prediction accuracy, achieving minimal values for MSE (0.0042) and PRD (0.8450).

KEYWORDS

ConvLSTM, MapReduce, MSE, PRD, S-SGD.

DOI: 10.9781/ijimai.2024.01.003

I. INTRODUCTION

AGRICULTURE is the broadest economic sector that has contributed to the socio-economic development of India. The economy and climate are the primary factors having the major impact on agriculture [1]. Crop production is enhanced using advanced agricultural techniques, which assist the monitoring system and provide information regarding the environment. Crop yield information is essential for industrial operation as the product of agriculture is used in industries. Decisions regarding the supply chain are planned by estimating the crop production accurately. Enhancing the crop yield is as important as monitoring the environmental factors. Agriculture depends on rainfall, cultivation, irrigation, fertilizers, pesticide weeds, soil, climate, temperature, harvesting, and other factors [2]. The growth of crops becomes a critical challenge in the absence of adequate rainfall.

It is important to know how much water is needed to get the right amount of water for the irrigation process [3].

The Indian Meteorological Department (IMD) divides the climatological seasons into four categories: summer or pre-monsoon, winter, monsoon, or rainy season, and fall or post-monsoon. Considering December to February, the winter season is in full swing. From March to May, the summer of pre-monsoon season begins. We experience the rainy season, sometimes known as the monsoons from June through September. The southwest summer monsoon, which is humid generally starts during late May or early June, and it moves slowly. The monsoon rains begin to fade at the starting period, especially in the early October. South India gets more rainfall than the rest of the nation. The post-monsoon season begins in October–November. The Northeast Monsoon is when Tamil Nadu receives most of its yearly rainfall. The agricultural sector predominantly depends on

Please cite this article in press as: S. O. Manoj, A. Kumar, A. K. Dubey, J. P Ananth. An Adaptive Salp-Stochastic-Gradient-Descent-Based Convolutional LSTM With MapReduce Framework for the Prediction of Rainfall, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 4, pp. 32-44, 2025, <http://dx.doi.org/10.9781/ijimai.2024.01.003>

the monsoon. The forecasting of the amount of rain in a specific area is a tedious process. It requires careful observation of the atmospheric conditions and the development of models that stimulate cumulus cloud interaction and atmospheric conditions, leading to a high degree of complexity [4].

In addition, rainfall prediction is even more difficult. Estimating the rainfall comprises investigating the groundwater quality, soil features, land features, and other aspects. Misleading conclusions occur if any of the factors are neglected without any consideration. For example, when estimating the required amount of water for agriculture, if the amount of rainfall in the particular area does not account for and the water is more than predicted for the specific area, then water-logging occurs, resulting in salinity, which, in turn, reduces the crop yield.

On the other hand, if the water is less than indicated, it leads to under-irrigation, which results in reduced yield efficiency. As a result, predicting the appropriate quantity of rainfall is of paramount importance for the irrigation of crops [5]. Physical and data-driven models [6],[7] are the two types of models that are used for predicting rainfall. Physical models use physical principles to simulate the physical processes that contribute to the rainfall process, whereas data-driven models forecast the future based on historical data [8].

Agriculture primarily depends on the soil condition and climatic features. There is a significant need for a system or model that can reliably forecast precipitation. That's why we're working on it in the first place: to create a rain forecasting system that uses deep learning and the MapReduce framework. This approach contributes by passing the data into an adaptive salp-stochastic-gradient-descent-based convolutional long short-term memory (adaptive-S-SGD-based ConvLSTM) network. The optimal weights have been carefully chosen by the adaptive S-SGD algorithm.

Following are the challenges found in the approaches corresponding to the prediction of rainfall confront:

- Analysis of precipitation using non-linear historical time series data is time-consuming [9].
- The dynamic nature of the weather is one of the challenges that need to be considered during weather forecasting as the weather data varies with time and space. The computation of the regression coefficients based on the past models during the training period and the observed forecast that neglects the dynamic nature of the weather is required [9], [10].
- The deviation in the spatial and the temporal rainfall conditions, the inaccurate initial conditions, and the complexity related to the physical process also pose a challenge for the prediction [11].
- Even though there are many tools for weather forecasting, predicting the weather data with a large structure and volume is a massive task. Thus, the prediction of weather using weather data is not an easy task [12].

The following are the most important contributions:

- The goal of this work is twofold: (1) to develop an adaptive S-SGD-based ConvLSTM system for predicting rainfall, and (2) to disseminate this knowledge to farmers so that they may maximise their crop yields.
- To demonstrate the superior performance of the proposed model, we have compared the results of the adaptive S-SGD-based ConvLSTM model to those of the S-SGD-based ConvLSTM model, ConvLSTM, and CLR.

In this study, we have used an adaptive S-SGD-based ConvLSTM system for rainfall prediction. The adaptive prediction model, built within the mappers of the MapReduce framework, is given to be the input data. The deep LSTM can be trained optimally by considering the optimal weights obtained using the adaptive S-SGD algorithm.

The algorithm is evolved by hybridizing the adaptive concept in the S-SGD algorithm. The hybridized algorithm is applied to ConvLSTM to adopt the best training in the LSTM system.

A serious concern around the world is rainfall prediction. It has attracted the attention of industries, the government, the research, and development sector, and also entities related to risk monitoring and management. Precipitation forecasting using an adaptive neuro fuzzy inference system-genetic algorithm (ANFIS-GA) was developed by Wahyuni et al. [13]. Although the ANFIS-GA model predicted the rainfall accurately, this technique failed to determine the crop budding period. Hussain et al. [14] designed a dynamic, self-organized multilayer neural network model to forecast naturally occurring signals. They used this method for clustering by changing the output layer with the backpropagation algorithm. However, this method required global optimization algorithms to achieve accurate predictions. Bhomia et al. [15] modeled a dynamical-model-selection-based multimodal ensemble model to predict the amount of rain during the rainy season. This method removed the non-normality of the predictors using transfer functions. However, the prediction rates need to be improved. Haidar and Verma [16] developed a deep convolutional neural network to predict rainfall. Although this method performed better with high annual averages, the diversities in the models need to be combined with the ensemble techniques. Zhou et al. [17] tried the deep learning method for predicting the rainfall for diverse types of convective climate. Poornima and Puspahatha [18] proposed a recurrent neural network (RNN) based on intensified LSTM towards rainfall prediction. This work's limitation is that the accuracy improvement is small compared to the existing LSTM and RNN methods. Anh et al. [19] proposed a method based on a deep learning model incorporated with LSTM and the feedforward neural network (FFNN) for precipitation downscaling for regions in Vietnam. Diop et al. [20] investigated using an innovative hybrid method, namely multilayer perceptron-Wale optimization algorithm, to predict the annual rainfall and used three lags. Khan and Maity [21] proposed a hybrid deep learning approach based on a multilayer perceptron (MLP) and the grouping of a one-dimensional convolutional neural network for multi-step-ahead daily rainfall prediction. Their approach combines an MLP and the support vector regression method. Zhang et al. [22] suggested a short-term rainfall forecasting model based on MLP. Putra et al. [23] presented a deep autoencoder-based semi-convolutional neural network for yearly rainfall prediction. Hewage et al. [24] suggested a lightweight model named a data-driven weather-predicting model by exploring the temporal modeling approaches of LSTM and temporal convolutional networks. They compared their performance with machine learning and ensemble methods. Neural networks have been combined with many other models to achieve accurate prediction power. The variation in the predicting power was to use the unique features of each model to identify the different patterns in the data. A severe restriction of using neural networks is the non-interpretation of the data analysis model. The conventional fuzzy systems may not have any learning procedure to shape the analysis model. The general circulation model (GCM) does not provide an accurate portrayal of the local climate. In addition to this, issues also exist in the various rainfall prediction techniques in agriculture. Many machine learning models use the gradient descent algorithm to tune and select the parameters. SGD is used in the fast computation of the requirements. The SGD algorithm might not achieve accuracy, but it fluctuates around the region close to the global minimum. When combined with the techniques such as swarm optimization, it provides optimized results. Artificial-neural-network (ANN)-based methods have been extensively used for predicting rainfall in various regions of the world. It is observed that RNNs such as LSTM enhance the explicit management of order between the observations whenever the knowledge of a mapping function is obtained from the inputs to the outputs. This LSTM provides native support for the

sequences. To do defensible time forecasting, methods such as LSTMs [25] and hybrid LSTMs can be used. Ridwan et al. [26] suggested a rainfall forecasting model using machine learning. The prediction obtained using their model did not monitor the water level in the reservoirs. Methods such as Bayesian linear regression, boosted decision tree regression, and neural network regression were used in the process. Here, the daily error rate and the monthly error rate were calculated. The primary focus of the authors was to introduce the RAIN-F+ dataset, and only the early fusion methods were used. The results thus obtained were validated only for the radar observations. Jaseena et al. [27] executed a deep survey on the techniques based on deep learning to process massive datasets for rainfall prediction. Their study covered the idea that only small datasets were used in most of the works and that massive datasets needed to be utilized. The evaluation of stability was not addressed in this work. Ahmed et al. [28] aggregated the antecedent lag memory of the climate node indices and the rainfall. A hybrid LSTM method was used in the process and was influenced by the low-frequency variability of the climate indices. This type of forecast was applied for flood forecasting and irrigation scheduling. Sun et al. [29] proposed a convolutional 3D gated recurrent unit model to predict rainfall intensity over a short period of time based on machine learning. The limitation of this work is that it has shortcomings in the prediction of rainfall within an hour. Tran et al. [30] suggested an improved rainfall prediction model using the combined preprocessing methods and FFNN. It provided better results compared to the conventional models but exhibited poor results compared to the hybrid models. Lin et al. [31] proposed a hybrid deep learning algorithm and its application to streamflow prediction. Their model comprised the first-order difference, FFNN, and LSTM. The limitation of this work is the short lead time. Velasco et al. [32] suggested week-ahead rainfall forecasting using the MLP neural networks. In this work, the hidden neurons are limited, and thus, the performance is observed to be minimal compared to the existing methods. Dewitte et al. [33] proposed methodologies related to weather forecasts based on artificial intelligence techniques. The model gave better results when machine learning and deep learning features were incorporated. Hussein et al. [34] demonstrated the importance of using well-grounded statistical techniques and compared their analysis results with various predictive models based on machine learning. The limitation of this work is that the authors have used only a single parameter prediction that is based on the spatiotemporal features, and the applicability of their model is limited in most of the cases. Yan et al. [35] proposed a rainfall forecast model based on the TabNet model. The reliability of the model was good, and if the addition of more data and parameter tuning had been done, the robustness could have been increased. Convolutional LSTM plays a significant role in the time series data analysis. Deep neural networks can accurately predict rainfall. The salp swarm algorithm (SSA) can be used as one of the better prediction algorithms. The hybrid algorithms provide better forecasts than the individual algorithms. Thus, we have incorporated the SSA and the SGD algorithms [34] based on a convolutional neural network to predict rainfall that will support the farming community and agriculture. In 2018, Yang and Yang [36] proposed, a modified CNN algorithm with dropout and SGD optimizer to enhance feature recognition, reduce CNN time-cost. Achieved high recognition rates on MNIST, HCL2000, and EnglishHand datasets, outperforming other methods. Table I shows the literature analysis of the current trends.

From the analysis of the existing works, we can conclude that the traditional approaches take a long time to process, limited data is used and needs more computational requirements. Neural Network has been used in combination with many other models to give a better prediction. The goal of combining these forecasting models was to take advantage of each one's strengths to predict future events more accurately. One major drawback of Neural Networks is that the

underlying data analysis model might not understand the results. The analysis model used by traditional fuzzy systems cannot be learned. General circulation models (GCM) fail to accurately portray the regional climate. Also covered are the difficulties encountered by various agricultural rainfall prediction methods. Gradient Descent is widely used in ML models for parameter selection and tuning. When a rapid computation is essential, SGD is used. While the SGD may not be able to pinpoint an exact value, it does tend to float around the neighbourhood of the global minimum. When combined with the techniques like Swarm optimization, it provides optimized results.

The ANN based techniques has been used for a long period of time to predict the rainfall in various regions of the world. It is observed that, the Recurrent Neural networks like the LSTM adds explicit handling of order between the observations, whenever the learning of a mapping function happens from inputs to outputs. This LSTM gives the native support for the sequences. To do defensible time forecasting, methods like LSTMs and Hybrid LSTMs can be used. Convolutional LSTM plays a major role in the times series data analysis. Deep Neural Networks are likely to be used in the prediction of rainfall and is proved to give a better prediction. The Salp Swarm Optimization Algorithm can be used as one of the better prediction algorithms. The hybrid algorithms have given better predictions than the individual algorithms. So, we have incorporated the SSA along with the Stochastic Gradient Descent algorithm based on a convolutional neural network for the prediction of rainfall that will be a support to the farming community and agriculture. Deep Learning, when applied to Machine Learning, offers insightful replacements for analysing massive data sets and is useful for autonomous feature selection. To analyse data in real time and generate reliable findings and analysis, Deep Learning has developed fast and efficient algorithms and data-driven models. The ability to foretell precipitation using a variety of methods developed over time. Precipitation forecasting is where Deep Learning algorithms really shine.

What follows is the rest of the paper's outline: The ConvLSTM technique based on S-SGD has been addressed in Section II. Discussion and analysis of the results are presented in Part III. In part IV, we draw a final conclusion.

II. METHODS

For planning agriculture, rainfall prediction plays a major role. The existing rainfall prediction methods faced issues while dealing with big data. Hence, in this research, the MapReduce framework is used to overcome the issues related to big data, such as parallel computing and computational complexity. The MapReduce framework used the adaptive S-SGD-based ConvLSTM networks for the prediction of the rainfall. The data regarding the weather is given as the input to the framework, which contains both the map and reducer functions. Here, weather data is considered as time-series data and its large amount of data is given as input for dealing with the rainfall prediction process. The MapReduce framework can be used to solve the complexities associated with big data in an efficient manner. It also has the capability to perform parallel computing processes.

A. Adaptive S-SGD Method Towards Training the ConvLSTM Network

1. ConvLSTM Model

Because of its ability to highlight what is most obviously present in the field of view, CNN is commonly employed in feature engineering. LSTM has the characteristic of increasing in size with the passage of time, making it a useful tool for time series analysis. With the capabilities of CNN and LSTM in mind, we create a model for predicting rainstorms using CNN with an LSTM. The main

TABLE I. LITERATURE SURVEY OF THE METHODOLOGIES, APPROACH, AND LIMITATION

Research	Objectives	Methods used	Approach	Dataset	Result	Limitation	Advantage
(Janarthananet al. 2021) [37]	The primary objective of this research is to support people by informing them about the unlikeable natural catastrophes.	Fuzzy logic is used to build effective applications or fuzzy base systems to continuously control and monitor the automatic stream engine. Mainly, qualitative, and imprecise expressions have been used in this research to predict rainfall using a fuzzy base system.	Classical controller approaches that are also known as a "point to point control system" have been used in the research to control and range the system.	The US Department of Agriculture (USDA) scan data and rainfall data is used in this research.	As per the dataset, the rainfall volume, like temperature, wind speed, and humidity, is high resulting in heavy rainfall has taken place.	Often, predictions (temperature estimation) may not be accurate because it uses fuzzy datasets.	By predicting the forecast using low-cost FLC, it is possible to support humans and reduce the danger. Therefore, people will be able to receive relevant information about natural incidents in advance.
(Haq et al. 2021) [38]	The main objective of this research is to predict rainfall to expect danger in advance.	Machine learning methods such as Support Vector Machine (SVM), Artificial Neural Network (ANN), and mean absolute error are used in this research study.	LSTM and deep learning approach has been used to predict daily temperature and rainfall.	Indian Ocean Dipole and El Nino Index data 3.4 is used in this research to enables researchers to predict the certain dangers and rainfall patterns simultaneously.	The prediction using IOD and Nino parameters enables people to identify the rainfall pattern on the 6th week.	Mean arctangent absolute percentage error (MAAPE) can calculate only a limited range of values to predict the actual error.	The El-Nino and IOD characteristics were strong enough to predict rainfall in the Sidoarjo area of East Java, according to the MAAPE prediction findings, with a value of 0.9644.
(Pham et al., 2020) [39]	The main objective of this research is to build and compare several AI models, such as FIS and adaptive network.	Critical success index (CSI), Probability of detection (POD), Mean absolute error (MAE), and correlation coefficient	Monte Carlo, MLP, decision tree, and K- Nearest Neighbour (KNN) that enables researchers to analyse which model is effective for rainfall predictions.	Training and testing dataset with 3653 sample data.	From January to April, SVM and ANN models have predicted a small volume of rainfall	Accurate prediction of rainfall	There is a chance of predicting the accuracy of rainfall with an appropriate AI model.
(Zhao et al. 2021) [40]	The main objective of this research is to build an AI-based prediction model for calculating the flow of debris.	Model evaluation method, Tsfresh rolling technique, data processing methods, and ensemble methods (gradient descent and extra tree).	Machine learning approach	Cost validation dataset	The researcher has been able to make a good balance between missing alarms and false alarms. Furthermore, this AI-based prediction model is able to reduce the timing of false alarms.	The limited sample size is used to target 70% of random data for training purposes. Otherwise, it can take more time to train more data.	Early warning on the flow of debris and receive accurate rainfall and threshold prediction model.
(Shrestha et al. 2020) [41]	The primary objective is to analyze the use of intelligent computing in robotics.	Kernel method, Network embedding method, and SVM	Parallel random forest and MAV with a low-cost approach	Standard image dataset and GRU network dataset	The robot can monitor the person's activity and their shopping actions.	Calculation in server related index, fingerprint scan through the offline and limited control device.	Data collection through offline mode is possible.
(Chhetri et al. 2020) [42]	The primary objective is the pattern recognition technique for rainfall prediction.	Numeric weather prediction and deep learning methods	Pattern recognition and gated recurrent unit approach	Weather dataset of the period of 1979 to 2009	MLP provides better results about rainfall prediction as compared to the short-term model	Limited computational resources	Improvement of the future prediction model can be possible with the MLP technique.
(Diez-Sierra and del Jesus, 2020) [43]	The primary objective is to predict the long-term pattern of rainfall by utilizing an atmospheric synoptic pattern.	Machine learning and statistical methods	SVM, KNN, k-means, and random forest	Rainfall dataset for managing agency and water planning of Tenerife Island	The f-score metric is beneficial for identifying the accuracy of each model while working with an unbalanced dataset.	Limited sample size	The specific application of the PCA method enables people to receive original and accurate predicted data.
(Zhang et al., 2020) [44]	The main objective is to propose a Tiny RainNet by joining CNN along with "bi-directional long short-term memory."	Optical flow method, machine learning method (SVM), and radar quantitative rainfall prediction (QPF)	Machine learning and big data algorithm	CIKM AnalytiCup 2017	Tiny RainNet requires only 3 ms time to check the whole dataset	Limited data in a single mapRader	This tiny RainNet model provides accurate results about rainfall prediction within a minimum time.

Research	Objectives	Methods used	Approach	Dataset	Result	Limitation	Advantage
(Li et al. 2021) [45]	The primary objective of this research is to develop a data-driven model of food predicting the streamflow with precipitation.	Gridded surface subsurface hydrologic and LSTM	Feature selection approach	US National Elevation, validation, and calibration datasets	The result section discussed that the LSTM model is able to predict rainfall patterns with a good performance level.	Limitation about the computational expense	Accurate prediction of rainfall with LSTM network.
(Hewage et al. 2021) [46]	The main objective of this research is to forecast accurate weather in a reliable NWP that includes WRF and met office models.	Dynamic ensemble method, persistent precision method, and traditional cross-correlation method.	Statistical (quantitative) approach and machine learning approach (feedforward neural network)	Training, testing, and validation dataset.	The result section exhibits that ANN is more reliable in terms of weather forecasting than the cross-validation method.	Ineffective design methodologies for specifying parameters and massive computational requirements.	Advancement of technology and identification of accurate models for weather forecasting is possible in the future.
(Venkatesh et al. 2021) [47]	The primary objective of this research is to develop a rainfall prediction system by utilizing an adversarial network to predict the rainfall data and future rainfall of India.	Financial method, finite difference method, deep learning method, and GAN based method.	ANN and regression approach.	A real rainfall dataset is used in this research study to validate the rainfall prediction system effectively.	The experimental results exhibit that the proposed system offers 99% of accurately predicted results of rainfall.	This proposed system is unable to distinguish the generated data of rainfall from the original data.	The designed GAN model can generate higher predicted rainfall values than the real rainfall data.
(Ren et al. 2021) [48]	The main objective of this research is to predict accurate and timely weather.	Numerical weather prediction (NWP), deep learning, and data-driven methods.	Deep learning-based weather prediction (DLWP) and a data-driven approach are used in this research.	Observation dataset that includes situation analysis data, remote sensing data, and simulation analysis data of the NWP model.	The result section reveals that weather forecasting results vary based on climate change.	Massive computational requirements	The timeliness and accuracy of DWLP are better than the NWP that, supports people in analysing the weather pattern in advance.
(Kumar et al. 2021) [49]	The primary objective of this research is to develop a deep learning-enabled downscaling model for analysing India summer monsoon rainfall data.	Super-resolution convolutional neural network (SRCNN), SVM, and other deep neural methods.	Researchers have tracked traditional SR approach, ML-based approach, and correlation coefficient approach to inspect the summer monsoon rainfall pattern.	IMD gridded dataset	The result section reveals that high-resolution data that is derived from deep learning models provides accurate results rather than linear interpolation.	Limited in spatial resolution	Accurate information based on India's summer monsoon rainfall data can be inspected.
(Fayaz et al. 2022) [50]	The main objective of this work to develop the rainfall prediction method in the worst-case weather scenarios that will be anticipated in advance, and appropriate warnings will be delivered.	Methods from both the nonlinear autoregressive with exogenous input (NARX) neural network and the adaptive grey wolf levenberg-marquardt (GWLm) network were combined.	Trained using the grey wolf optimizer (GWO) and the Levenberg-Marquardt (LM) algorithms	past weather data of Kashmir province, India	The MSE and R-values of the suggested model were examined, and the value of error should be minimized for successful outcomes, as shown in the results section. In addition, performance parameters, including specificity, sensitivity, specificity, recall, accuracy, and Cohen kappa are assessed.	Takes a long time to construct these models and look into the different datasets.	Reaches the best accuracy measures
(Rahman et al. 2022) [51]	novel real-time rainfall prediction system for smart cities using a machine learning fusion technique.	Naive Bayes, K-nearest Neighbors, Support Vector Machines, and Decision Trees are employed. For accurate precipitation forecasting, the framework employs fuzzy logic, also known as fusion, to combine the results of various machine learning techniques.	Machine learning techniques with fuzzy logic	12 years of historical weather data (2005 to 2017) for the city of Lahore	The proposed work is compared with different techniques in terms of Accuracy rate and Miss rate.	The data which will be used for prediction is compromised	higher accuracy

components of this model are the input layer, the one-dimensional convolution layer, the pooling layer, the long short-term memory (LSTM) hidden layer, and the fully connected layer. The complete approach is shown in Fig. 1. All the components of Fig. 1 have been discussed subsequently in this section.

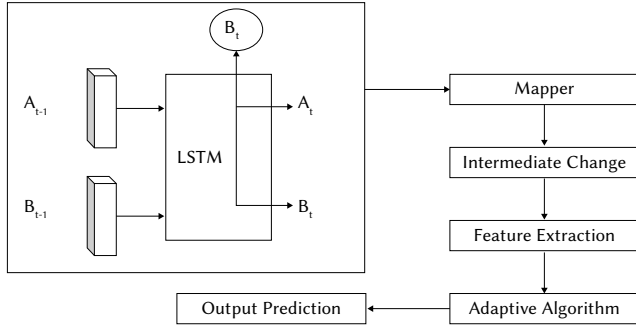


Fig. 1. Computational aspect of the proposed approach.

Here, the ConvLSTM network architecture that effectively handles the time-series data for rainfall prediction is described. This network has inputs as A_1, \dots, A_t , hidden states B_1, \dots, B_t , along with the gates C_t, D_t, E_t and the outputs F_1, \dots, F_t . The future is predicted using the past states associated with the neighbors and the inputs that are assisted by the two operators, namely convolutional operator, which is represented as (\bullet) and the Hadamard product, that is represented as $(*)$. Effective prediction is provided by organizing the ConvLSTM network into forecasting and encoding layers, especially when we employ a large transitional kernel. The LSTM network consists of memory units that have cells and gates. The gates and the memory cells control the information flow. The forecast is assured based on the spatiotemporal sequences of ConvLSTM. The new input is updated in the state of the memory cell of the ConvLSTM, the unnecessary contents are forgotten, and the outputs are obtained. The input gate and output of the memory unit are computed as follows: Equation (1) describes the output at the input gate.

$$C_t = \gamma(\beta_C^A \bullet A_t + \beta_C^B \bullet B_{t-1} + \beta_C^F \bullet F_{t-1} + \sigma^C) \quad (1)$$

where the input vector is A_t , the gate activation function is represented as γ , β_C^F is the weight among the input layer and the cell output then we have the weight among the input layer and the memory output layer that is given as β_C^B , F_{t-1} and B_{t-1} are the output of the previous memory unit and cell, respectively, β_C^A is considered as the weight among the input gate and the input layer the bias of the input layer is given as σ^C , \bullet and \bullet are the multiplication operator which is considered element-wise and the convolutional operator, respectively. Forget gate D_t 's result is expressed as Equation (2).

$$D_t = \gamma(\beta_D^A \bullet A_t + \beta_D^B \bullet B_{t-1} + \beta_D^F \bullet F_{t-1} + \beta^D) \quad (2)$$

where the weight amongst the forget gate mode and the input layer is β_D^A , weight between the cell and the output gate is β_D^F , the weight between the memory unit of the earlier layer and the output gate is β_D^B , and the bias of the forget gate is β^D . The output at the output gate is given by Equation (3).

$$E_t = \gamma(\beta_E^A \bullet A_t + \beta_E^B \bullet B_{t-1} + \beta_E^F \bullet F_t + \sigma^E) \quad (3)$$

where $\beta_E^A, \beta_E^B, \beta_E^F$ are the weights that link the input layer present and the output gate, the memory unit and the output gate, and the cell and the output gate, respectively. σ^E is the bias at the output gate. The activation function of the weights is determined as the output of the temporary cell and is expressed as Equation (4).

$$\tilde{F}_t = \tan H(\beta_H^A \bullet I_t + \beta_H^B \bullet B_{t-1} + \sigma^H) \quad (4)$$

where β_H^A is the weight among the cell and the input layer, β_H^B is considered as the weight among the cell and the memory unit, and σ^H is the bias at the cell. The sum of the difference between the memory unit of the earlier layer and the current layer and the temporary cell state is the output of the cell. The output is given by Equation (5) and Equation (6).

$$F_t = R_H \bullet F_{t-1} + C_t \bullet \tilde{F}_t \quad (5)$$

$$XF_t = XD_t \bullet XF_{t-1} + XC_t \bullet \tan H(\beta_H^A \bullet XA_t + \beta_H^B \bullet XB_{t-1} + \sigma^H) \quad (6)$$

The memory unit's output is given by Equation (7).

$$B_t = E_t \bullet \tan H(F_t) \quad (7)$$

where E_t is the output gate and B_t is represented as the memory block output. The output layer's output is given as Equation (8).

$$P_t = \psi(\beta_P^B \bullet B_t + \sigma^P) \quad (8)$$

where σ^P is the bias of the output layer, P_t is considered as the output vector, β_P^B is the weight between the memory unit and the output layer. The bias and the weight of the ConvLSTM are given by $\beta \in \{\beta_P^B, \beta_H^A, \beta_H^B, \beta_E^A, \beta_E^B, \beta_E^F, \beta_D^A, \beta_D^B, \beta_D^F, \beta_C^A, \beta_C^B, \beta_C^F\}$ and $\sigma \in \{\sigma^H, \sigma^D, \sigma^E, \sigma^C\}$, respectively. For the final prediction, the output that is received from the LSTM's encoding layer is served into the forecasting layer. The proposed S-SGD algorithm optimally tunes the weight and the biases of ConvLSTM.

2. Training Algorithm for ConvLSTM

The weights obtained through the ConvLSTM are trained by the adaptive S-SGD algorithm, which is the hybridization of the adaptive concept in the S-SGD algorithm, whereas the S-SGD algorithm is the alteration of the available SGD algorithm using the methodology in the SSA algorithm. The modification of the SGD algorithm with SSA helps in enhancing the ideal universal convergence capability of the algorithm [52]. The algorithmic processes of the proposed adaptive S-SGD algorithm are given as follows:

Process a: Initialization Step: At the initial stage, we have to initialize the weights which are represented by the solution vector, $k(t)$. At a certain time, t , $kd(t)$; ($1 \leq d \leq e$) is the representation of the solution of the algorithm

Process b: Objective Function Evaluation: The second step is where the objective function is evaluated. Here, the classifier is trained for getting the optimal solution. The optimal function is given by Equation (9).

$$\text{Min}(k) = \frac{V}{2} \|k\|^2 + fn(i_{m_t}, j_{m_t}) \quad (9)$$

The minimum value of the objective function is selected as the optimal solution.

Process c: Updating the weights using the S-SGD algorithm: The proposed algorithm is obtained by adapting the SGD algorithm along with SSA. The proposed algorithm updates the weights by altering the S-SGD algorithm with the adaptive concept. The SGD is mathematically expressed Equation (10).

$$k_d(t) = \left(\frac{t}{t-1}\right) [k_d(t+1) - i_{m_t} \cdot j_{m_t}] \quad (10)$$

Where i_{m_t} the feature is vector and j_{m_t} is the m^{th} training sample. The weights at iteration $(t+1)$ are given by $k_d(t+1)$ and $k_d(t)$, respectively. Based on the training input and the weights of the previous iteration, the standard update equation is computed. The chosen sample will have a target at a time t and is represented as m_q and i_{m_t}, j_{m_t} are the m^{th} training samples. Among the training data, the randomly chosen training sample is given by (i_{m_t}, j_{m_t}) . The SSA update

equation is given by Equation (11).

$$k_d(t+1) = \frac{1}{2}[k_d(t) + k_{d-1}(t)] \quad (11)$$

Where $k_d(t+1)$ is the d^{th} salp position at iteration $(t+1)$ is the iteration. The SGD is modified by replacing the updated equation of SSA in the SGD, which is given by Equation (12).

$$k_d(t+1) = \frac{2(t-1)}{t-2} \left\{ \frac{1}{2} \times k_{d-1}(t) - \frac{1}{2} \times \left(\frac{t}{t-1} \right) \times i_{m_t} \cdot j_{m_t} \right\} \quad (12)$$

where $k_d(t+1)$ is the newly updated weight, $k_{d-1}(t)$ is considered as the weight at each iteration which is evolved from the previous iteration, and (i_{m_t}, j_{m_t}) is the sample of training. The classifier is trained for articulating the optimal weights using the Equation (12). For adaptive prediction of the optimal weights, the adaptive concept is included in Equation (13).

$$k_d(t+1) = \frac{t-1}{t-2} \left(k_{d-1}(t) - \frac{t}{t-1} \times i_{m_t} \cdot j_{m_t} \right) \quad (13)$$

Here, as Equation (13) depends on iteration; it can be considered as an evolutionary factor that can be taken from PSO. The Equation (13) then becomes Equation (14).

$$k_d(t+1) = f(k_{d-1}(t) - \frac{t}{t-1} \times i_{m_t} \cdot j_{m_t}) \quad (14)$$

Where $f = \frac{v_p - v_{min}}{v_{min} \max \in [0,1]}$. From Equation (14), the finest result is selected based on the objective function that is minimum, and the best solution is generated centered on the objective function of the preceding iteration and the training sample (Equation (15) and Equation (16)).

$$v_m = \frac{1}{2-1} \sum_{d=1}^2 \sqrt{\sum_{g=1}^C (k_d^g(t) - k_{d-1}^g(t))^2} \quad (15)$$

$$v_m = \sum_{d=1}^2 \sqrt{\sum_{g=1}^C (k_d^g(t) - k_{d-1}^g(t))^2} \quad (16)$$

Process d: Termination: Steps b and c are repetitive until the iteration reaches its maximum number.

An example of the pseudo code for the proposed adaptive S-SGD algorithm is shown in Algorithm 1, with the goal of deriving optimal weights via ConvLSTM network training.

Algorithm 1: Adaptive S-SGD algorithm

INPUT: Solution vector, $kd(t)$, $(1 \leq t \leq q)$

OUTPUT: Optimal weights, $kd(t+1)$.

Begin

Initialization

Population Initialization is done

Evaluation of Objective Function

Train the classifier using optimal function in Equation 9

Update the weights

Modify the S-SGD with adaptive concept as in Equation 10

SSA update as in Equation 11

Modify SGD by substituting update equation of SSA in SGD as in Equation 12

Adaptive prediction of optimal weights is included as in Equation 13

Take the evolutionary factor from PSO as in Equation 14

Best solution is generated on the objective function of the preceding iteration and the sample training data

Repeat the process

End

B. MapReduce Framework Along With Deep Learning Approach for Rainfall Prediction

The forecast of rainfall is one among the vital part in agriculture development and planning. Most rainfall prediction models utilizing weather data have failed to deal with the available massive amount of data that is called as big data [14]. As an example of huge data, consider the weather data, which can be time-series data. The MapReduce framework can solve the problems associated with big data, reducing the computational complexity and providing parallel computing. The input to the MapReduce framework is the weather data that uses the adaptive S-SGD-based ConvLSTM network and the map and reduce functions for an active rainfall prediction process. Using the suggested adaptive S-SGD algorithm, the weights of a ConvLSTM network are trained. This algorithm modifies the adaptive notion found in the S-SGD algorithm. Various delays are applied to the outputs of the individual mappers to train them. Finally, the output of the mapper is concatenated and provided as input to the reduction, which produces the final forecast. Fig. 2 depicts the suggested method for rainfall forecasting. Using the MapReduce framework, the input meteorological data is considered to be F and is utilised to predict precipitation.

1. Rainfall Prediction Based on the MapReduce Framework

The visualization is improved and the MapReduce framework provides an efficient prediction of rainfall. The framework's processing power is reduced by the given MapReduce framework by storing and distributing the datasets through a large number of servers. The ConvLSTM is trained by passing the weather data into the individual mappers. The input data is mapped using the mapper and reducer functions of MapReduce programming. These two functions of the framework based on the MapReduce model utilizes the adaptive S-SGD-based ConvLSTM network. Weather data, which acts as the input is processed using the number of mappers available in the mapper module. The mapper's output is used with intermediate data to train the reducer for the final prediction.

Mapper Phase: The first module, entitled mapper, is operated using the adaptive S-SGD algorithm that has been proposed. Consider the availability of a total of x mappers that are denoted as Equation (17).

$$Q = \{Q_1, Q_2, \dots, Q_n, \dots, Q_x\} \quad (17)$$

Where Q_n is the n^{th} mapper. The input weather data which has various delays are trained using the ConvLSTM in the individual mappers. For example, mapper-1 with the ConvLSTM is trained using $F(b - [D + 1])$ to yield the predicted output, $F(b + w)$. In the same way, mapper-2 is trained using $F(b - [D + 2])$ to give the predicted output as $F(b + w - 1)$. The weather data, $F(b)$, is used for training the x^{th} mapper for providing the predicted output of $F(b + 1)$. The prediction is performed by the individual mapper considering the previous records with a delay of $[D + 1]$, $[D + 2]$, etc. The output that is predicted at the mapper is denoted as,

$$\begin{aligned} F_{x+1}, \dots, F_{x+wt} = \\ \operatorname{argmax}_{F_{x+1}, \dots, F_{x+wt}} \omega(F_{wt-D+1}, F_{wt-D+2}, \dots, F_{wt}) \end{aligned} \quad (18)$$

$$\begin{aligned} F_{x+1}, \dots, F_{x+wt} \approx \\ \operatorname{argmax}_{F_{x+1}, \dots, F_{x+wt}} \omega(F_{x+1}, \dots, F_{x+wt} | R_{\text{encode}}(F_{x-D+1}, F_{x-D+2}, \dots, F_x)) \end{aligned} \quad (19)$$

$$F_{x+1}, \dots, F_{x+wt} \approx \eta_{\text{forecast}} \omega(R_{\text{encode}}(F_{x-D+1}, F_{x-D+2}, \dots, F_x)) \quad (20)$$

The mapper output trains the inbuilt function with ConvLSTM called reducers.

Reducer phase: ConvLSTM performs the final rainfall prediction

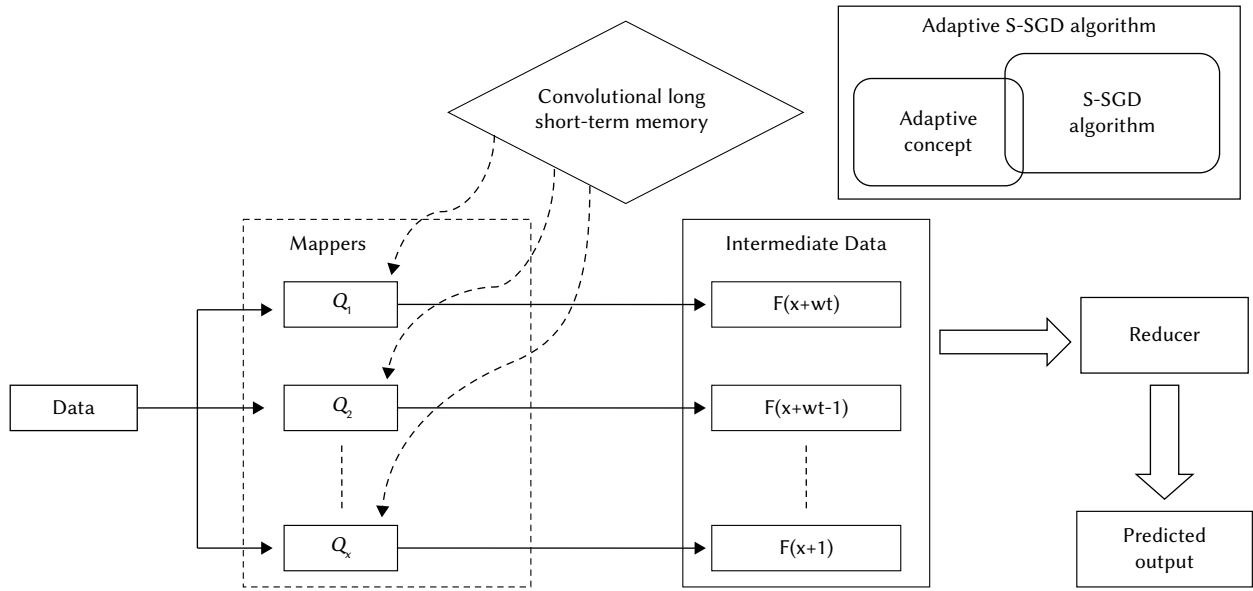


Fig. 2. Proposed rainfall prediction model's Block diagram.

with the reduce function. Using the suggested adaptive S-SGD technique, the ConvLSTM is optimally tuned during the reducer phase. The data input at the reducer is represented by Equation (21).

$$F^{inter} = \{F(b+w), F(b+w-1), \dots, F(b+1)\} \quad (21)$$

In order to get the final prediction, the intermediate data (F^{inter}) trains the reducer. The expected result is $F(b+1)$.

Testing phase: The testing phase is one of the critical phases in the process. During the testing step, the concatenated output is obtained by feeding test data to the mappers with varying delays. The reducer of the MapReduce framework then generates the precisely predicted results.

III. RESULTS AND DISCUSSION

This part discusses the findings obtained using the adaptive S-SGD-based ConvLSTM algorithm suggested in the previous section. To demonstrate the effectiveness of the suggested method, a comparison with existing methodologies has been conducted.

A. Experimental Setup

The recommended method was implemented in MATLAB software, and the rainfall forecast dataset was analysed [28]. The dataset contains subdivision-specific precipitation data from 1901 to 2015 spanning 115 years. The column-by-column data are presented as follows: The first twelve columns contain precipitation data from January to December. The following column has the annual precipitation data, followed by four columns containing quarterly precipitation data: JF (January and February), MAM (March, April, and May), JJAS (June, July, August, and September), and OND (October, November and December). The dataset was obtained from both data.gov.in and the Indian Meteorological Department in Pune.

These datasets were obtained via India's Open Government Data website. The dataset contains monthly rainfall information for Tamil Nadu districts. This dataset has been subdivided into three subsets containing information on annual, monthly, and quarterly precipitation.

In dataset 1, monthly rainfall series data were extracted from the dataset, including India's rainfall data. In dataset 2, annual rainfall series data were extracted from the dataset, including India's rainfall data. In dataset 3, quarterly rainfall series data were extracted from the

dataset, including India's rainfall data. The dataset 4 comprises rainfall data for Tamilnadu, as well as a sequence of rainfall data by month. The dataset 5 provides rainfall data for Tamilnadu that is organised by year, and the dataset 6 contains rainfall data for Tamilnadu that is organised by quarter.

Deep analysis was performed using these six datasets, and the results were compared based on the performance metrics.

B. Performance Metrics

The proposed adaptive S-SGD-based ConvLSTM method was analyzed using the performance metrics, such as the mean square error (MSE) and the percentage root mean square difference (PRD).

MSE: MSE is the mean square difference between the expected output and estimated output. The value of MSE should be small if the method is efficient. MSE is computed as follows Equation (22):

$$MSE = \frac{1}{S} \sum_{l=1}^S (R_l - R)^2 \quad (22)$$

where S is the total number of samples, R_l is the estimated output, and R is the target output.

PRD: PRD is used to assess how trustworthy a certain approach is in producing precise results. It is formulated as Equation (23).

$$PRD = \sqrt{\frac{\sum_{l=1}^S (R_l - R)^2}{\sum_{l=1}^S (R_l)^2}} \times 100 \quad (23)$$

The proposed adaptive S-SGD based ConvLSTM method was compared with the existing methods, namely, S-SGD-based ConvLSTM, ConvLSTM, and CLR. This section presents the results of a comparative investigation of different rainfall prediction systems, wherein the MSE and PRD performance metrics were calculated using the six datasets.

Using dataset 1: The comparative results obtained from the various rainfall prediction methods for different training data size using dataset 1 are shown in Fig. 3 and Fig. 4.

The MSE-based analysis is shown in Fig. 3. For a training data size of 0.8, the MSE produced by the proposed adaptive S-SGD based ConvLSTM method, the existing S-SGD based ConvLSTM, and the CLR are 0.00851, 0.00871, 0.01887, and 0.00871, respectively. When compared to other methods currently in use, it is clear that the MSE value of the suggested method is the lowest.

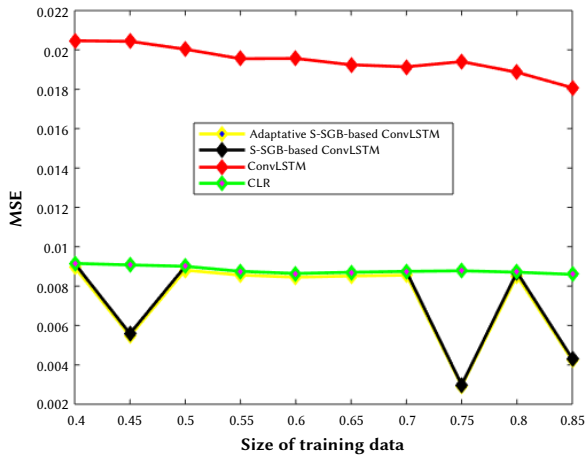


Fig. 3. MSE - Analysis using dataset 1.

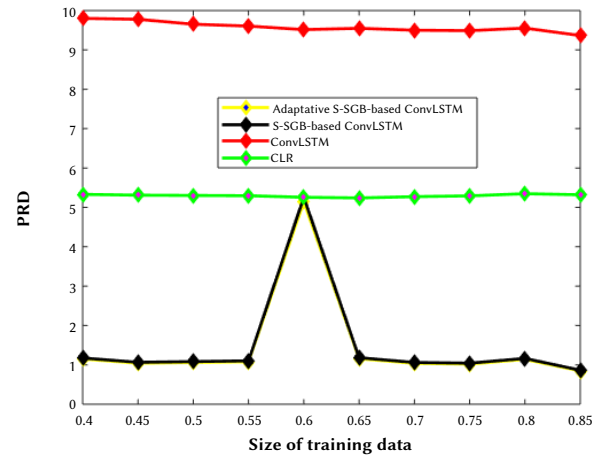


Fig. 4. PRD - Analysis using dataset 1.

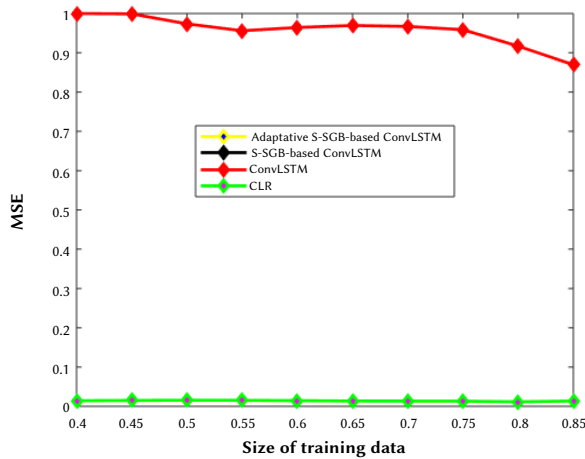


Fig. 5. MSE - Analysis using dataset 2.

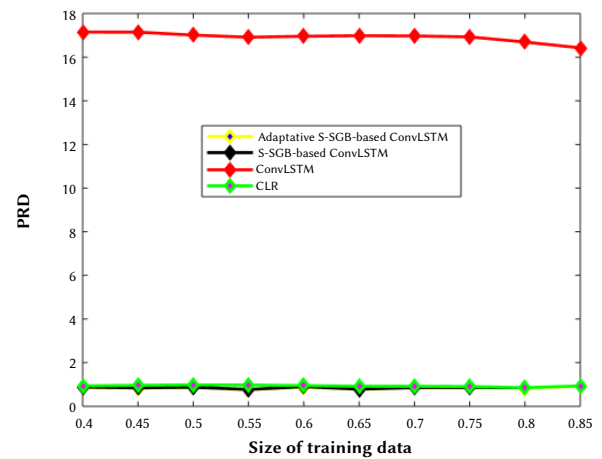


Fig. 6. PRD - Analysis using dataset 2.

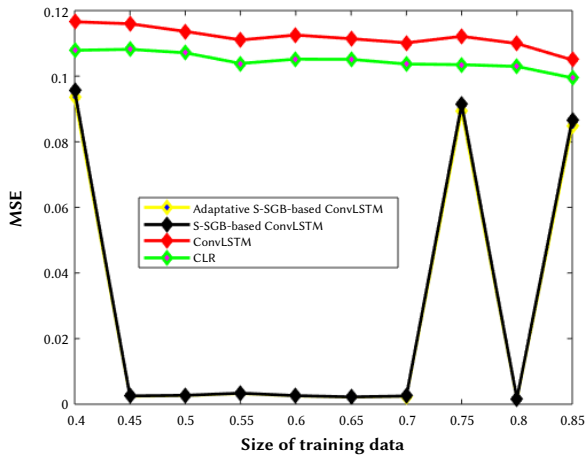


Fig. 7. MSE - Analysis using dataset 3.

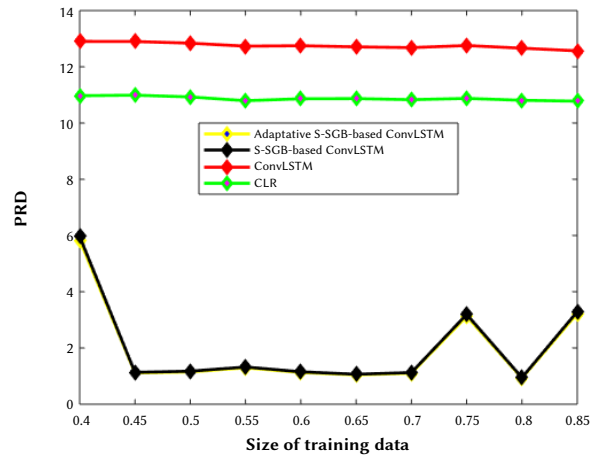


Fig. 8. PRD - Analysis using dataset 3.

Fig. 4 shows the comparative analysis based on PRD. Considering the training data size as 0.8, the PRD obtained by the proposed adaptive S-SGD based ConvLSTM and the existing S-SGD-based ConvLSTM, ConvLSTM, and CLR is 1.1377, 1.1633, 9.5547, and 5.3480, respectively.

Using dataset 2: Fig. 5 and Fig. 6 show the comparative results obtained from the various rainfall prediction methods for different sizes of training data using dataset 2. Prediction methods are assessed based on their MSE, as illustrated in Fig. 5. For a training data size of 0.8, the adaptive SSGD-based ConvLSTM achieves MSE values of

0.0112, 0.0115, 0.9169, and 0.0115, respectively.

The analysis of PRD-based prediction methods is presented in Fig. 6. For instance, at a training data size of 0.8, the proposed adaptive S-SGD-based ConvLSTM, the existing S-SGD-based ConvLSTM, and the ConvLSTM yield PRD values of 0.8340, 0.8527, 16.6927, and 0.8527, respectively. In comparison to other approaches, the suggested method demonstrates the lowest MSE and PRD values.

Using dataset 3: The comparative results obtained from the various rainfall prediction methods for different training data size using

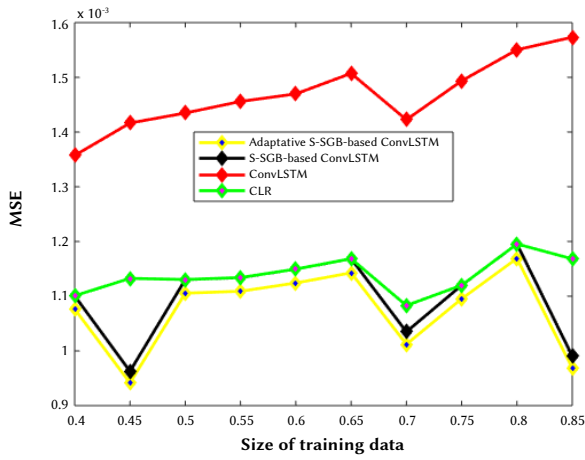


Fig. 9. MSE - Analysis using dataset 4.

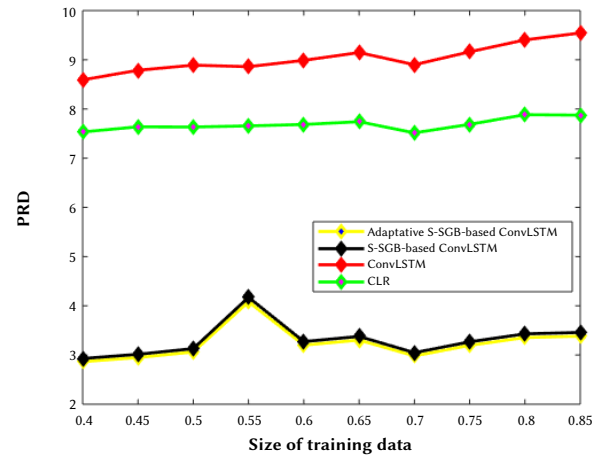


Fig. 10. PRD - Analysis using dataset 4.

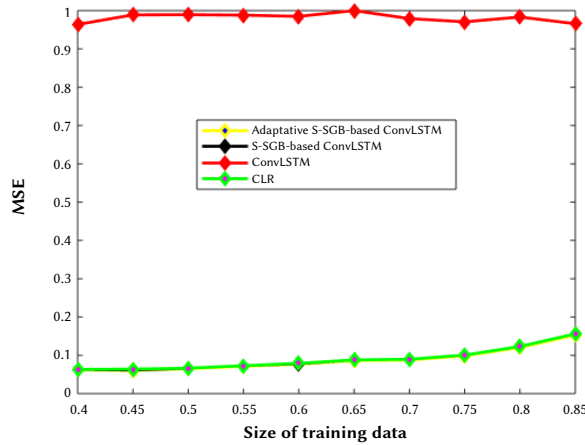


Fig. 11. MSE - Analysis using dataset 5.

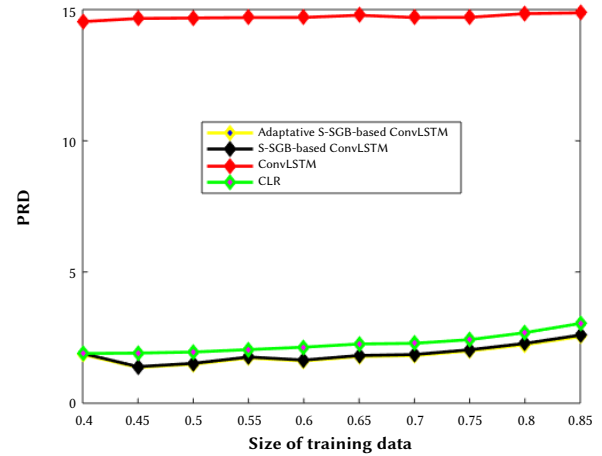


Fig. 12. PRD - Analysis using dataset 5 - PRD.

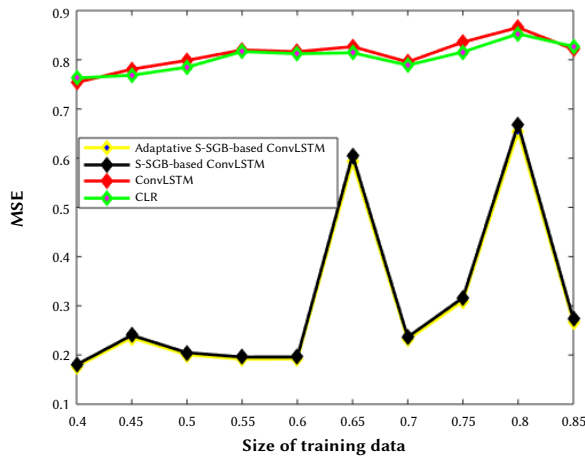


Fig. 13. MSE - Analysis using dataset 6.

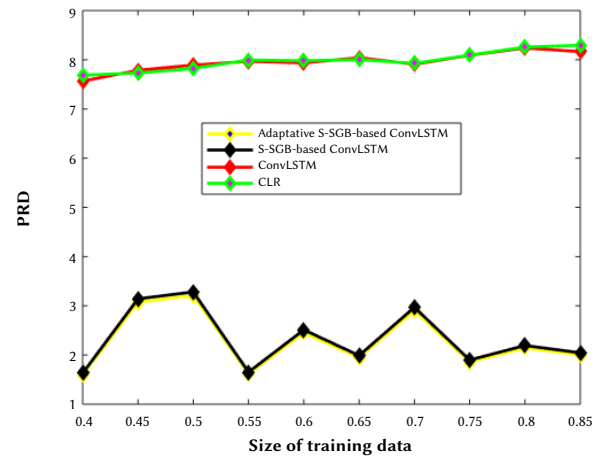


Fig. 14. PRD - Analysis using dataset 6.

dataset 3 are shown in Fig. 7 and Fig. 8.

Fig. 7 represents the analysis based on MSE. When the training data size is 0.85, the MSE obtained by the proposed adaptive S-SGD based ConvLSTM, and the existing S-SGD-based ConvLSTM, ConvLSTM, and CLR is 0.0848, 0.0867, 0.1050, and 0.0995, respectively. From the values, it is clear that the MSE value of the proposed method is the lowest compared to those corresponding to the other existing methods.

Fig. 8 depicts the comparative analysis based on PRD. Considering the training data size as 0.85, the PRD corresponding to the proposed adaptive S-SGD based ConvLSTM and the existing S-SGD-based

ConvLSTM, ConvLSTM, and CLR is 3.2199, 3.2923, 12.5674, and 10.7774, respectively.

Using dataset 4: Fig. 9 and Fig. 10 show the comparative results obtained from the various rainfall prediction methods for different size of training data using dataset 4.

Fig. 9 depicts the analysis of the prediction methods based on MSE. MSE obtained by the proposed adaptive S-SGD-based ConvLSTM and the existing S-SGD-based ConvLSTM, ConvLSTM, and CLR for a training data size of 0.85 is 0.00097, 0.00099, 0.00157, and 0.00117, respectively.

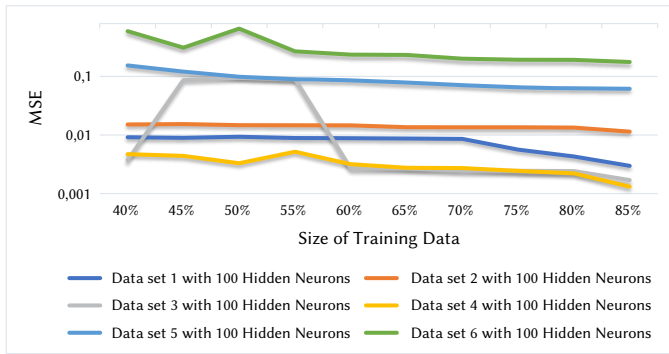


Fig. 15. Overall MSE comparison of the complete six datasets.

Fig. 10 depicts the analysis of the prediction methods based on PRD. PRD corresponding to the proposed adaptive S-SGD-based ConvLSTM and the existing S-SGD-based ConvLSTM, ConvLSTM, and CLR for a training data size of 0.85 is 3.3828, 3.4589, 9.5439, and 7.8748, respectively. The proposed method exhibits the lowest MSE and PRD values compared to the other existing methods.

Using dataset 5: The comparative analysis of the various rainfall prediction methods for different training data size using dataset 5 is shown in Fig. 11 and Fig. 12.

Fig. 11 depicts the analysis based on MSE. Considering the training data size to be 0.85, the MSE obtained by the proposed adaptive S-SGD-based ConvLSTM and the existing S-SGD-based ConvLSTM, ConvLSTM, and CLR is 0.15259, 0.15603, 0.96569, and 0.15604, respectively. The analysis shows that the MSE value of the proposed method is the lowest compared to the other existing methods.

Fig. 12 shows the comparative analysis based on PRD. Considering the training data size as 0.85, the PRD obtained by the proposed adaptive S-SGD-based ConvLSTM and the existing S-SGD-based ConvLSTM, ConvLSTM, and CLR is 2.5377, 2.5948, 14.8740, and 3.0397, respectively.

Using dataset 6: Fig. 13 and Fig. 14 show the comparative analysis of the rainfall prediction methods for different size of training data using dataset 6.

Fig. 13 depicts the analysis of the prediction methods based on MSE. MSE values corresponding to the proposed adaptive S-SGD-based ConvLSTM and the existing S-SGD-based ConvLSTM, ConvLSTM, and CLR for a training data size of 0.85 are 0.2673, 0.2733, 0.8215, and 0.8270, respectively.

Fig. 14 depicts the analysis of the prediction methods based on PRD. The PRD values corresponding to the proposed adaptive S-SGD-based ConvLSTM and the existing S-SGD-based ConvLSTM, ConvLSTM, and CLR for a training data size of 0.85 are 1.9996, 2.0446, 8.1656, and 8.2950, respectively. The proposed method exhibits the lowest MSE and PRD values compared to the other existing methods.

The performance of the proposed Adaptive SGD-based ConvLSTM as well as existing methods such as convLSTM CLR and S-SGD based ConvLSTM is evaluated using performance measures such as MSE and PRD. According to the overall performance study of adaptive S-SGD based ConvLSTM in terms of MSE for the 6 datasets, the adaptive S-SGD based model performs best for Tamilnadu monthly rainfall prediction, followed by India's quarterly rainfall prediction. According to the overall performance study of adaptive S-SGD based ConvLSTM in terms of PRD for the 6 datasets, the adaptive S-SGD based model performs best for India's yearly rainfall forecast, followed by India's monthly rainfall prediction.

The overall performance analysis of adaptive S-SGD based ConvLSTM in terms of MSE for the 6 datasets with 100 hidden neurons

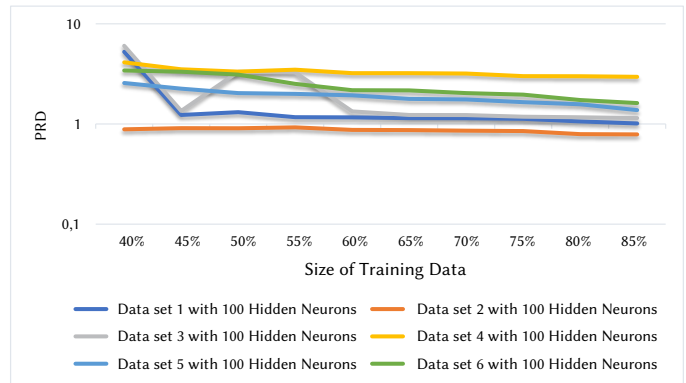


Fig. 16. Overall PRD comparison of the complete six datasets.

is observed, and it is stated that the adaptive S-SGD based model performs well for monthly rainfall prediction of Tamilnadu, followed by the quarterly rainfall prediction of India. The next priority is given to the monthly and yearly prediction of rainfall in India, followed by Tamilnadu's yearly and quarterly rainfall prediction. It has been shown in Fig. 15 and Fig. 16.

Some of the major findings of the proposed method for rainfall prediction in agriculture are illustrated below.

- The simulation results of the proposed Adaptive SGD-based ConvLSTM and SGD-based ConvLSTM are compared with the existing techniques, like ConvLSTM and CLR.
- The proposed Adaptive SGD-based ConvLSTM provides MSE as 0.00292 and PRD value as 0.84501.
- The proposed SGD-based ConvLSTM has values of 0.00298 and 0.86402 as MSE and PRD, respectively.
- Among all the comparative methods, the proposed Adaptive SGD-based ConvLSTM has improved performance for MSE and PRD.
- Thus, it is clearly shown that the proposed Adaptive SGD-based ConvLSTM considerably increases the system performance.
- The proposed algorithms can efficiently predict the monthly, quarterly, and yearly rainfall.

The main limitations of this study and the future suggestions are as follows:

- This study can be extended to discover other interesting patterns in the time series data analysis using optimization algorithms.
- In this paper, we have not considered rainfall prediction in a short span.
- The future studies may focus on getting the prediction accuracy without directly using the processed data for analysis.
- Other datasets need to be adopted to check the validity of schemes in real-time.

IV. CONCLUSION

As part of this effort, an adaptive SSA-based ConvLSTM system was created to reliably predict precipitation from a given set of meteorological inputs. For the weather forecast, we took in dynamically updated time series data processed with the MapReduce framework. Implementing the proposed adaptive S-SGD-based ConvLSTM model in the MapReduce framework proved highly helpful in dealing with the big data problems. The mapper and reducer components of this architecture are helpful in estimating precipitation. The suggested adaptive S-SGD-based ConvLSTM was fed the input data, and its weights were fine-tuned using a MapReduce

framework. The suggested rainfall prediction model was evaluated using six datasets taken from the Rainfall Prediction database. The MSE and PRD evaluation metrics were used for the performance study of the proposed approach. The proposed technique was shown to have better prediction accuracy, with an MSE and PRD of 0.0042 and 0.8450, respectively. To improve the accuracy of this technique, a hybrid fusion model for forecasting precipitation should be created. However, only two measures of efficiency were considered in this paper. Accuracy, Precision, F-measure, and Recall are just a few examples of future performance metrics that will be analysed.

REFERENCES

- [1] J. Majumdar, S. Naraseeyappa and S. Ankalaki, "Analysis of agriculture data using data mining techniques: application of big data," *Journal of Big data*, vol. 4, no. 1, pp. 1-5, 2017.
- [2] S. Rajeswari, K. Suthendran and K. Rajakumar, "A smart agricultural model by integrating IoT, mobile and cloud-based big data analytics," In *IEEE international conference on intelligent computing and control*, pp. 1-5, 2017.
- [3] I. Pani, D. D. Putranto and P. K. Wardhani, "Net present value (NPV) of the rehabilitated irrigation channels to increase agricultural production," *International Journal of Advanced Technology and Engineering Exploration*, vol. 8, no. 78, pp. 576-583, 2021.
- [4] A. Y. Barrera-Animas, L. O. Oyedele, M. Bilal, T. D. Akinosho, J. M. Delgado and L. A. Akanbi, "Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting," *Machine Learning with Applications*, vol. 7, 2022.
- [5] B. Abishek, R. Priyatharshini, M. A. Eswar and P. Deepika, "Prediction of effective rainfall and crop water needs using data mining techniques," In *IEEE technological innovations in ICT for agriculture and rural development*, pp. 231-235, 2017.
- [6] S. Jha, A. Dey, R. Kumar, V. Kumar, "A novel approach on visual question answering by parameter prediction using faster region based convolutional neural network," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 5, pp. 30-37, 2019.
- [7] P. C. Chiu, A. Selamat, O. Krejcar, K. K. Kuok, E. Herrera-Viedma, G. Fenza, "Imputation of Rainfall Data Using the Sine Cosine Function Fitting Neural Network," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 7, pp. 39-48, 2021.
- [8] A. M. Bagirov, A. Mahmood and A. Barton, "Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach," *Atmospheric Research*, vol. 188, pp. 20-29, 2017.
- [9] A. K. Dubey, A. Kumar, V. García-Díaz, A. K. Sharma and K. Kanhaiya, "Study and analysis of SARIMA and LSTM in forecasting time series data," *Sustainable Energy Technologies and Assessments*, vol. 47, 2021.
- [10] M. A. Rahman, L. Yunsheng and N. Sultana, "Analysis and prediction of rainfall trends over Bangladesh using Mann-Kendall, Spearman's rho tests and ARIMA model," *Meteorology and Atmospheric Physics*, vol. 129, no. 4, pp. 409-424, 2017.
- [11] A. Sharifian, Á. Fernández-Llamazares, H. Wario, Z. Molnár and M. Cabeza, "Dynamics of pastoral traditional ecological knowledge: a global state-of-the-art review," *Ecology and Society*, vol. 27, no. 1, pp. 1-63, 2022.
- [12] Z. Beheshti, M. Firouzi, S. M. Shamsuddin, M. Zibarzani and Z. Yusop, "A new rainfall forecasting model using the CAPSO algorithm and an artificial neural network," *Neural Computing and Applications*, vol. 27, no. 8, pp. 2551-2565, 2016.
- [13] I. Wahyuni, W. F. Mahmudy and A. Iriany, "Rainfall prediction using hybrid adaptive neuro fuzzy inference system (ANFIS) and genetic algorithm," *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 9, no. 2, pp. 51-56, 2017.
- [14] A. J. Hussain, P. Liatsis, M. Khalaf, H. Tawfik and H. Al-Asker, "A dynamic neural network architecture with immunology inspired optimization for weather data forecasting," *Big Data Research*, vol. 14, pp. 81-92, 2018.
- [15] S. Bhomia, N. Jaiswal, C. M. Kishtawal and R. Kumar, "Multimodel prediction of monsoon rain using dynamical model selection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 5, pp. 2911-2917, 2016.
- [16] A. Haidar and B. Verma, "Monthly rainfall forecasting using one-dimensional deep convolutional neural network," *IEEE Access*, vol. 6, pp. 69053-69063, 2018.
- [17] K. Zhou, Y. Zheng, B. Li, W. Dong and X. Zhang, "Forecasting different types of convective weather: A deep learning approach," *Journal of Meteorological Research*, vol. 33, no. 5, pp. 797-809, 2019.
- [18] S. Poornima and M. Pushpalatha, "Prediction of rainfall using intensified LSTM based recurrent neural network with weighted linear units," *Atmosphere*, vol. 10, no. 11, 2019.
- [19] D. Tran Anh, S. P. Van, T.D. Dang and L.P. Hoang, "Downscaling rainfall using deep learning long short-term memory and feedforward neural network," *International Journal of Climatology*, vol. 39, no. 10, pp. 4170-4188, 2019.
- [20] L. Diop, S. Samadianfard, A. Bodian, Z. M. Yaseen, M. A. Ghorbani and H. Salimi, "Annual rainfall forecasting using hybrid artificial intelligence model: integration of multilayer perceptron with whale optimization algorithm," *Water Resources Management*, vol. 34, no. 2, pp. 733-746, 2020.
- [21] M. I. Khan and R. Maity, "Hybrid deep learning approach for multi-step-ahead daily rainfall prediction using GCM simulations," *IEEE Access*, vol. 8, pp. 52774-52784, 2020.
- [22] X. Zhang, S. N. Mohanty, A. K. Parida, S. K. Pani, B. Dong and X. Cheng, "Annual and non-monsoon rainfall prediction modelling using SVR-MLP: an empirical study from Odisha," *IEEE Access*, vol. 8, pp. 30223-30233, 2020.
- [23] A. B. Putra, R. Malani, B. Supratty and A. F. Gaffar, "A Deep Auto Encoder Semi Convolution Neural Network for Yearly Rainfall Prediction," In *IEEE International Seminar on Intelligent Technology and Its Applications*, pp. 205-210, 2020.
- [24] P. Hewage, A. Behera, M. Trovati, E. Pereira, M. Ghahremani, F. Palmieri and Y. Liu, "Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station," *Soft Computing*, vol. 24, no. 21, pp. 16453-16482, 2020.
- [25] D. Zhang, G. Lindholm and H. Ratnaweera, "Use long short-term memory to enhance Internet of Things for combined sewer overflow monitoring," *Journal of Hydrology*, vol. 556, pp. 409-418, 2018.
- [26] W. M. Ridwan, M. Sapitang, A. Aziz, K. F. Kushiar, A. N. Ahmed and A. El-Shafie, "Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia," *Ain Shams Engineering Journal*, vol. 12, no. 2, pp. 1651-1663, 2021.
- [27] K. U. Jaseena and B. C. Kovoor, "Deterministic weather forecasting models based on intelligent predictors: A survey," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 3393-3412, 2020.
- [28] A. M. Ahmed, R. C. Deo, Q. Feng A. Ghahramani, N. Raj, Z. Yin and L. Yang, "Deep learning hybrid model with Boruta-Random Forest optimiser algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity," *Journal of Hydrology*, vol. 599, 126350, 2021.
- [29] D. Sun, J. Wu, H. Huang, R. Wang, F. Liang and H. Xinhua, "Prediction of short-time rainfall based on deep learning," *Mathematical Problems in Engineering*, 2021.
- [30] D. Tran Anh, T. Duc Dang and S. Pham Van, "Improved rainfall prediction using combined pre-processing methods and feed-forward neural networks," *J*, vol. 2, no. 1, pp. 65-83, 2019.
- [31] Y. Lin, D. Wang, G. Wang, J. Qiu, K. Long, Y. Du, H. Xie, Z. Wei, W. Shangguan and Y. Dai, "A hybrid deep learning algorithm and its application to streamflow prediction," *Journal of Hydrology*, vol. 601, 126636, 2021.
- [32] L. C. Velasco, R. P. Serquiña, M. S. Zamad, B. F. Juanico and J. C. Lomoco, "Week-ahead rainfall forecasting using multilayer perceptron neural network," *Procedia Computer Science*, vol. 161, pp. 386-397, 2019.
- [33] S. Dewitte, J. P. Cornelis, R. Müller and A. Munteanu, "Artificial intelligence revolutionises weather forecast, climate monitoring and decadal prediction," *Remote Sensing*, vol. 13, no. 16, 2021.
- [34] E. A. Hussein, M. Ghaziasgar, C. Thron, M. Vaccari and Bagula, "Basic statistical estimation outperforms machine learning in monthly prediction of seasonal climatic parameters," *Atmosphere*, vol. 12, no. 5, 2021.
- [35] J. Yan, T. Xu, Y. Yu and H. Xu, "Rainfall forecast model based on the tabnet model," *Water*, vol. 13, no. 9, 2021.

- [36] J. Yang and G. Yang, "Modified convolutional neural network based on dropout and the stochastic gradient descent optimizer," *Algorithms*, vol. 11, no. 3, 2018.
- [37] R. Janarthanan, R. Balamurali, A. Annapoorani and V. Vimala, "Prediction of rainfall using fuzzy logic," *Materials Today: Proceedings*, vol. 37, pp. 959-963, 2021.
- [38] D. Z. Haq, D. C. Novitasari, A. Hamid, N. Ulinnuha, Y. Farida, R. D. Nugraheni, R. Nariswari, H. Rohayani, R. Pramulya and A. Widjayanto, "Long short-term memory algorithm for rainfall prediction based on El-Nino and IOD data," *Procedia Computer Science*, pp. 829-837, 2021.
- [39] B. T. Pham, L. M. Le, T. T. Le, K.T Bui, V. M. Le, H. B. Ly and I. Prakash, "Development of advanced artificial intelligence models for daily rainfall prediction," *Atmospheric Research*, vol. 237, 104845, 2020.
- [40] Y. Zhao, X. Meng, T. Qi, Y. Li, G. Chen, D. Yue and F. Qing, "AI-based rainfall prediction model for debris flows," *Engineering Geology*, vol. 296, 106456, 2022.
- [41] M. Shrestha, S. P. Panday, B. Joshi, A. Shakya and R. K. Pandey, "Automatic pose estimation of micro unmanned aerial vehicle for autonomous landing," In international conference on intelligent computing, pp. 3-15, 2020.
- [42] M. Chhetri, S. Kumar, P. R. Pratim and B. G. Kim, "Deep BLSTM-GRU model for monthly rainfall prediction: A case study of Simtokha, Bhutan," *Remote sensing*, vol. 12, no. 19, 2020.
- [43] J. Diez-Sierra, M. J. Del, "Long-term rainfall prediction using atmospheric synoptic patterns in semi-arid climates with statistical and machine learning methods," *Journal of Hydrology*, vol. 586, 124789, 2020.
- [44] C. J. Zhang, H. Y. Wang, J. Zeng, L. M. Ma and L. Guan, "Tiny-RainNet: a deep convolutional neural network with bi-directional long short-term memory model for short-term rainfall prediction," *Meteorological Applications*, vol. 27, no. 5, 2020.
- [45] W. Li, A. Kiaghadi and C. Dawson, "High temporal resolution rainfall-runoff modeling using long-short-term-memory (LSTM) networks," *Neural Computing and Applications*, vol. 33, no. 4, pp. 1261-1278, 2021.
- [46] P. Hewage, M. Trovati, E. Pereira and A. Behera, "Deep learning-based effective fine-grained weather forecasting model," *Pattern Analysis and Applications*, vol. 24, no. 1, pp. 343-366, 2021.
- [47] R. Venkatesh, C. Balasubramanian and M. Kaliappan, "Rainfall prediction using generative adversarial networks with convolution neural network," *Soft Computing*, vol. 25, no. 6, pp. 4725-4738, 2021.
- [48] X. Ren, X. Li, K. Ren, J. Song, Z. Xu, K. Deng and X. Wang, "Deep learning-based weather prediction: a survey," *Big Data Research*, 2021.
- [49] B. Kumar, R. Chattopadhyay, M. Singh, N. Chaudhari, K. Kodari and A. Barve, "Deep learning-based downscaling of summer monsoon rainfall data over Indian region," *Theoretical and Applied Climatology*, vol. 143, no. 3, pp. 1145-1156, 2021.
- [50] S. A. Fayaz, M. Zaman, and M. A. Butt, "A hybrid adaptive grey wolf Levenberg-Marquardt (GWLM) and nonlinear autoregressive with exogenous input (NARX) neural network model for the prediction of rainfall," *International Journal of Advanced Technology and Engineering Exploration*, vol. 9, no. 89, pp. 509-522, 2022.
- [51] A. U. Rahman, S. Abbas, M. Gollapalli, R. Ahmed, S. Aftab, M. Ahmad, M. A. Khan and A. Mosavi, "Rainfall Prediction System Using Machine Learning Fusion for Smart Cities," *Sensors*, vol. 22, no. 9, 2022.
- [52] J. P. Ananth and S. O. Manoj, "MapReduce and optimized deep network for rainfall prediction in agriculture," *The Computer Journal*, vol. 63, no. 6, pp. 900-912, 2020.



S. Oswalt Manoj

Dr. S. Oswalt Manoj is currently working as an Associate Professor in the Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Coimbatore, Tamil Nadu. He obtained his B.E degree in Computer Science and Engineering from Cape Institute of Technology, Tirunelveli in 2008 and his M.E Degree in Computer Science and Engineering from

University Departments, Anna University of Technology, Tirunelveli in 2010. He completed his PhD in the field of Information and Communication Engineering from Anna University Chennai with the title "Analysis and Prediction of Rainfall in Agriculture based on Deep Learning Framework". He has 13+ years of experience in Teaching, Research, Mentoring and Administration.

His research focuses on Data Mining, Machine Learning, Deep Learning, Big Data Analytics and Cloud computing. He has has memberships in more than 14 professional bodies. He has his publications in more than 30 International Journals, 25 International Conferences and 21 National Conferences. He has given guest lectures in recent trends for various Universities. Oswalt Manoj is also a reviewer and editorial board member for numerous journals and professional associations. He has more than 5 patents. He has acted as editor for various books.



Abhishek Kumar

Dr. Abhishek Kumar is Doctorate in computer science from University of Madras and done M.tech in Computer Sci. & Engineering from Government engineering college Ajmer, Rajasthan Technical University, Kota India. He has total Academic teaching experience of more than 10 years with more than 150 publications in reputed, peer reviewed National and International Journals, books & Conferences.

He has guided more than 20 M.Tech Projects and Thesis and Supervised 12 MS and 3 PhD scholars. His research area includes- Artificial intelligence, Image processing, Computer Vision, Data Mining, Machine Learning. He has been Session chair and keynote Speaker of many International conferences, webinars in India and abroad. He has been the reviewer for IEEE and Inderscience Journal. He has authored/Co-Author 7 books published internationally and edited 35 books (Published & ongoing with Elsevier, Wiley, IGI GLOBAL Springer, Apple Academic Press, De-Grueter and CRC etc. He has been member of various National and International professional societies in the field of engineering & research like Senior Member of IEEE, IAENG (International Association of Engineers), Associate Member of IRED (Institute of Research Engineers and Doctors), He has got Sir CV Raman National award for 2018 in young researcher and faculty Category from IJRP Group. He is Book Series Editor with Elsevier and Degruyter.



Ashutosh Kumar Dubey

Dr. Ashutosh Kumar Dubey is an Associate Professor in the Department of Computer Science at Chitkara University School of Engineering and Technology, situated in Himachal Pradesh, India. He is a Postdoctoral Fellow at the Ingenium Research Group Lab, Universidad de Castilla-La Mancha, Ciudad Real, Spain. Dr. Dubey earned his BE and M.Tech degrees in Computer Science and Engineering from RGPV, Bhopal, Madhya Pradesh. He completed his PhD in Computer Science and Engineering at JK LakshmiPat University in Jaipur, Rajasthan, India. He is a Senior Member of both IEEE and ACM and possesses more than 16 years of teaching experience. Dr. Dubey has authored and edited fifteen books and has published over 60 articles in peer-reviewed international journals and conference proceedings. He serves as an Editor, Editorial Board Member, and Reviewer for numerous peer-reviewed journals. His research interests encompass Machine Learning, Renewable Energy, Health Informatics, Nature-Inspired Algorithms, Cloud Computing, and Big Data.



J. P. Ananth

Dr. J. P. Ananth received his B.E (2000) and M.E (2005) in Computer Science and Engineering from M S University and Ph.D. (2012) from Sathyabama University, Chennai. He is a Senior member of IEEE and a member of IEEE Computer Society. Having 20+ years of teaching experience, presently he is working as a Professor in the Department of Computer Science and Engineering, Sri

Krishna College of Engineering and Technology, Coimbatore. His research interests include Computer Vision, Pattern Recognition, Artificial Intelligence and Data Analytics. His work has been documented in many journals including IET Image Processing, The Computer Journal, Wireless Networks Springer. He serves as a reviewer for several International Conferences and Journals including IEEE Access and a member of technical and executive committees for International Conferences.

A Robust Framework for Speech Emotion Recognition Using Attention Based Convolutional Peephole LSTM

Ramya Paramasivam¹, K. Lavanya², Parameshchhari Bidare Divakarachari^{3*}, David Camacho⁴

¹ Department of Computer Science and Engineering, Mahendra Engineering College (Autonomous), Mallasamudram (India)

² Department of Electronics and Communication Engineering, Velammal Engineering College, Chennai (India)

³ Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bengaluru (India)

⁴ Departamento de Sistemas Informáticos, Universidad Politécnica de Madrid, Calle Alan Turing s/n, Madrid, 28031, Madrid (Spain)

* Corresponding author: paramesh@nmit.ac.in

Received 13 May 2024 | Accepted 27 September 2024 | Early Access 5 February 2025



ABSTRACT

Speech Emotion Recognition (SER) plays an important role in emotional computing which is widely utilized in various applications related to medical, entertainment and so on. The emotional understanding improvises the user machine interaction with a better responsive nature. The issues faced during SER are existence of relevant features and increased complexity while analyzing of huge datasets. Therefore, this research introduces a well-organized framework by introducing Improved Jellyfish Optimization Algorithm (IJOA) for feature selection, and classification is performed using Convolutional Peephole Long Short-Term Memory (CP-LSTM) with attention mechanism. The raw data acquisition takes place using five datasets namely, EMO-DB, IEMOCAP, RAVDESS, Surrey Audio-Visual Expressed Emotion (SAVEE) and Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D). The undesired partitions are removed from the audio signal during pre-processing and fed into phase of feature extraction using IJOA. Finally, CP-LSTM with attention mechanisms is used for emotion classification. As the final stage, classification takes place using CP-LSTM with attention mechanisms. Experimental outcome clearly shows that the proposed CP-LSTM with attention mechanism is more efficient than existing DNN-DHO, DH-AS, D-CNN, CEOAS methods in terms of accuracy. The classification accuracy of the proposed CP-LSTM with attention mechanism for EMO-DB, IEMOCAP, RAVDESS and SAVEE datasets are 99.59%, 99.88%, 99.54% and 98.89%, which is comparably higher than other existing techniques.

KEYWORDS

Attention Mechanism, Convolutional Peephole Long Short-Term Memory, Feature Selection, Improved Jellyfish Optimization Algorithm, Speech Emotion Recognition.

DOI: 10.9781/ijimai.2025.02.002

I. INTRODUCTION

THE advent in the era of artificial Intelligence has attracted a greater number of researches to work on human-computer interaction [1]. The affective computing plays a significant role in interaction among human and the computer which is endowed in computers with the ability to observe and exhibit the emotion of the humans [2]. In general, the emotional state of the humans is evaluated based on the speech, body language and their facial expression. Among these, speech is a kind of natural method utilized for human communications which is comprised with linguistic and paralinguistic information [3]-[5]. The information related to context and language is present in the linguistic information whereas the paralinguistic information has information of gender, age, emotions and some more unique attributes [6]. Several researches reveals that the audio signal acts as a simple mean to execute the link among human and computers which in turn become familiar with human voice and helps to predict the emotion

[7]. An affective Speech Emotion Recognition (SER) is characterized by features on the basis of speech signals such as bandwidth, duration and frequency. Automatic approach involved in SER helps in various real time applications based on recognizing and detecting the mental and emotional state of individuals [8], [9]. SER is vastly utilized in real time applications such as human-computer interaction, call centers, healthcare and automated translation systems.

Speech emotion recognition has received a great deal of attention in Psychology and cognitive science, but data science has recently contributed significantly to the advancement of SER by highlighting a particularly captivating and motivating feature of human-machine interaction in voice communication. Further, it can be applied to the field of e-learning, automobile board systems, autonomous remote call centers, and student emotions recognition during lectures. This motivates the researchers to work on a few well-known voice computation and classification techniques to extract sentiments from

Please cite this article as: R. Paramasivam, K. Lavanya, P. B. Divakarachari, D. Camacho. A Robust Framework for Speech Emotion Recognition Using Attention Based Convolutional Peephole LSTM, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 4, pp. 45-58, 2025, <http://dx.doi.org/10.9781/ijimai.2025.02.002>

audio inputs, utilizing deep learning techniques like audio signal preprocessing, feature extraction and selection approaches, and eventually determining the accuracy of the suitable classifier. The process of recognizing the human emotion is complex process due to the dependency on various factors such as speaker, gender, age and dialect. In every individual stage of SER, the data processing, feature extraction, feature selection and classification plays a major role [10]-[12]. The stage of pre-processing is based on normalizing the signals, removal of noise and artifacts. The feature extraction helps to mine out the salient features of emotion using different feature extraction techniques. Next to this, feature selection takes place which has a great role in reducing the complexities of SER and finally, the classification is performed with the help of machine learning/deep learning techniques [13], [14]. Additionally, most of the datasets based on speech emotion are comprised only with utterance level label classes that hold the information of the emotion [15]-[17]. The researches have utilized different machine learning and deep learning techniques for an effective SER, but the usage of deep learning architectures exhibits better results in modelling emotions [18], [19]. The usage of deep learning techniques provides promising results, having the ability to define the low-quality attributes to higher attributes. Moreover, they have tendency to handle the unlabeled data, manage with complex speech attributes and processing large datasets [20]. By considering this, this research focuses on introducing an optimization-based feature selection approach and an effective recognition of emotion using a novel deep learning approach.

A. Contribution

The major contribution of this research study is listed as follows:

1. The acquisitioned raw data is pre-processed and fed into the stage of feature extraction where the prosodic features and the acoustic features are extracted.
2. The feature selection performed using the proposed Improved Jellyfish Optimization Algorithm (IJOA) by introducing sine and cosine factors at stage of exploration and premature convergence strategy is utilized in the stage of exploitation.
3. Finally, the SER takes place with the help of the Convolutional Peephole Long Short-Term Memory (CP-LSTM) with attention mechanism. The combination of the proposed CP-LSTM with attention mechanism helps in an effective SER by recognizing the signal pattern and results in better classification accuracy.

The remainder of the manuscript is structured in the following manner: Section II presents details about the recent research based on SER and Section III presents the overall process involved in the proposed framework while recognizing the human emotions; Section IV presents the experimental outcome while evaluating CP-LSTM and finally, overall conclusion of this research is described in Section V of the manuscript.

II. RELATED WORK

In this section, the recent researches based on speech emotion recognition are discussed along with their advantages and drawbacks.

Yildirim et al. [21] introduced a modified feature selection approach for speech emotion recognition. This research utilized cuckoo search algorithm along with the non-dominated sorting genetic algorithm. The generation phase of the initial population is modified and the feature selection is performed by maximizing the classifier accuracy and minimizing the number of features using binary cuckoo and non-dominated sorting genetic algorithm. At last, the results are evaluated based on machine learning techniques. However, performing an exhaustive search for all subset space is infeasible.

Agarwal and Om [22] introduced the optimized Deep Neural Network (DNN) for speech emotion recognition. The speech signals are de-noised using Adaptive Wavelet Transform with Modified Galactic Swarm Optimization (AWT-MGSO). The de-noised output is provided to the phase extracting the features and the feature selection is performed using Adaptive Sunflower Optimization (ASFO). Finally, classification is performed using DNN-Deer Hunting Optimization (DHO). However, implementing a greater number of optimization techniques results in computational complexity.

Manohar and Logashanmugam [23] developed a hybrid deep learning approach along with feature selection for recognizing the emotion of speech using Deer Hunting with Adaptive Search (DH-AS). The acquisitioned data is pre-processed by median filtering and artifact removal, then it is subjected to the stage of feature extraction. After this, the selection of optimal features takes place using DH-AS and the classification is performed using hybrid DNN and Recurrent Neural Network (RNN). However, the suggested approach was suitable only for the balanced dataset and was complex due to its recurrent nature.

Mustaqeem and Kwon [24] developed an optimal feature selection approach using two stream Deep Convolutional Neural Network (D-CNN). The spectrum and the spectrogram of the speech signals are considered, then the high-level discriminative features are obtained using 2D and 1D CNN. The Iterative Neighborhood Component Analysis (INCA) was utilized in the process of selecting the optimal features by removing the redundant information. Finally, the classification was performed using a softmax layer. However, the improper pre-processing leads to discrepancies and redundancies that diminish the efficiency of the overall model.

Chattopadhyay et al. [25] introduced a hybrid feature selection approach using Clustering based Equilibrium Optimizer and Atom Search Optimization (CEOAS) for recognizing emotions from speech signals. The features such as Linear Prediction Coding (LPC) and Linear Predictive Cepstral Co-efficient (LPCC) were extracted from audio signals. At last, the results are evaluated with two classifiers like K-Nearest neighbor and Support Vector Machine (SVM). The suggested approach diminishes the feature dimension and helps to enhance the classification ability. However, CEOAS exhibits premature convergence at the time of iterative searching process.

Kanwal et al. [26] developed a Density based Spatial Clustering with Noise Genetic Algorithm (DGA) to recognize the type of emotion from the speech. The obtained data is pre-processed by removing the unvoiced audio segments and the optimization of feature was performed using DGA. After this, the reduction of features takes place using Principal Component Analysis (PCA) and finally, the classification was performed using Support Vector Machine (SVM). However, the PCA was computationally imprecise for huge datasets.

Barsainyan and Singh [27] introduced optimized speech emotion recognition using 1D CNN. The obtained data is augmented using glottal inverse filtering, silent elimination and noise addition. After this, the feature selection based on spectral contrast, zero crossing rate and amplitude energy are considered. Finally, the classification was performed using normalized CNN and XGB algorithm. However, the suggested framework was not vulnerable to higher datasets due to limited model training.

Sun et al. [28] introduced a speech emotion recognition approach using Improved Masking based Empirical Model Decomposition and Convolutional Recurrent Neural Network (IMEMD-CRNN). Initially, the decomposition of speech was performed using IMEMD, which is based on disturbance assisted EMD that determines the nature of signals. After this, the 43-dimensional time frequency features are used to characterize emotion and the acquired features are fed into CRNN to recognize the emotions. However, the mode mixing occurs which diminishes the capability of IMEMD to decompose the signal.

Ottoni et al. [29] introduced a deep learning approach based on CNN and Long Short-Term Memory (LSTM) which is used to recognize the speech emotions based on meta learning approach. After the stage of data acquisition, the optimizers such as Adam optimizer, Stochastic Gradient and Adagrad were used to select the optimal learning rate for the dataset. Then, data augmentation is performed to enhance the diversity of audio samples. The augmented output is provided to the phase of extracting features and the extracted features are fed into classification which is performed using CNN-LSTM. However, the lack of feature selection leads to computational complexity and diminishes the efficiency of classification.

Nhat Truong Pham et al. [30] proposed Hybrid Data Augmentation (HDA) and Modified Attention-Based Dilated Convolutional and Recurrent Neural Network (mADCRNN) methods for speech emotion recognition. The mADCRNN model learns and extracts utterance-level features from 3D log MelSpec low-level data by combining dilated CNNs and dilated LSTM models with an attention mechanism. While dilated recurrent neural networks solve complex dependencies and the vanishing and inflating gradient problems, dilated CNNs gain wider receptive fields. In addition, the loss functions are rearranged to identify different emotional states by merging the SoftMax loss and the center-based losses. However, the number of layers and parameters in the suggested mADCRNN model made it complicated.

Zengzhao Chen et al. [31] demonstrated a parallel network for multi-scale SER based on connection attention mechanism (AMSNet) for speech emotion recognition. In the meantime, AMSNet enhances feature characterization and feature enrichment by using several speech emotion feature extraction modules based on the temporal and spatial properties of speech signals. Varying types of features are given varying weight values by the network fusion connection attention technique. The model's capacity to recognize emotions has increased as a result of the integration of different features through the use of weight values. However, because of their less conspicuous characteristics, neutral emotions are less often recognized. Since the characteristics of neutral emotions were not sufficiently evident and it was challenging to categorize the corresponding voice signal, this problem also occurred here.

Mustaqeem Khan et al. [32] demonstrated a Multimodal Speech Emotion Recognition (MSER) system to effectively identify the speech emotions. The suggested model makes use of both text and audio to accurately predict the emotion label. The suggested model uses CNN to process the text and raw speech signal before feeding the results to appropriate encoders for the extraction of semantic and discriminative features. In order to improve text and auditory cues interaction, the cross-attention mechanism has been implemented to both elements. This allows crossway to extract the most pertinent information for emotion recognition. Eventually, the deep feature fusion technique allows for interaction between various layers and routes by merging the region-wise weights from both encoders. Yet, in order to address the restriction of the complicated interaction between speech signals and transcripts, these investigations disregarded the limitations of past knowledge. Francesco Ardan Dal Ri et al. [33] demonstrated an extensive validation using CNN for speech emotion recognition. Here, the suggested CNN integrated with a Convolutional Attention Block was tested in a sequence of experimentations including a collection of four datasets namely RAVDESS, TESS, CREMA-D, and IEMOCAP. After analyzing the datasets, they executed a cross-validation among emotional classes belonging to every specified dataset, through the purpose to examine the generalization capabilities of extracted features. Apart from the accuracy improvement once it was trained, the minimum accuracies only attained by testing validates the individuality of the individual models on the feature extraction.

Heuristic Multimodal Real-Time Emotion Recognition (HMR-TER)

approach has been developed for enhancing e-learning, which was introduced by Du, Y. et al. [34]. By promptly offering feedback based on learners' facial expressions and vocal intonations, the e-learning was enhanced. This approach uses gesture recognition analysis to enhance participation and interaction and hybrid validation dynamic analysis to solve the low learner motivation. On the other hand, a few tests with a limited range of accuracy were made available. An Audio-Visual Automatic Speech Recognition (AV-ASR) system was proposed by S. Debnath et al. [35] to enhance the educational experience of those with physical disabilities by enabling hands-free computing. For visual speech data, the Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) and Grey-Level Co-occurrence Matrix (GLCM) are suggested. The results of the study demonstrate that the suggested system obtains 96.50% accuracy for audio speech recognition and 76.60% accuracy for visual speech. However, as cluster size grows, accuracy decreases. This is due to the fact that a larger cluster size exhibits a dispersed representation of the data, lowering accuracy.

An Automatic Speech Recognition (ASR) system for the Lithuanian language was reported by L. Pipiras et al. [36]. It depends on deep learning techniques and recognizes spoken words only based on their phoneme sequences. The ASR task is solved using two different encoder-decoder models: a conventional model and a model with an attention mechanism. These models' effectiveness has been evaluated in two tasks: extended phrase recognition and isolated voice recognition. With shorter input sequences, the proposed model works merely well; however, it struggles with larger sequences. Bidirectional Long Short-Term Memory (BiLSTM) neural network and Wavelet Scattering Transform with Support Vector Machine (WST-SVM) classifier were developed by A. Lauraitis et al. [37] to identify patients' speech impairments at the beginning of central nervous system disorders (CNSD). The voice recorder from the Neural Impairment Test Suite (NITS) has been employed to capture speech data. Pitch contours, auditory spectrograms, Mel-frequency cepstral coefficients (MFCC), Gammatone cepstral coefficients (GTCC), and Gabor (analytic Morlet) wavelets are the sources of features that are extracted. Although there is a significant association among phoneme and grapheme sequence lengths in the corpus of data, not all patterns have the same length.

Table I shows the summary of literature reviews based on speech emotion recognition.

A. Problem Definition

The collected literature works show that the overall outcome of the existing approaches and the recognition ability to classify speech got affected due to the inappropriate techniques for selecting the relevant features; further, the complexity faced by the model to evaluate the huge datasets and poor classification methods relies as the reason to diminish the classification accuracy. Moreover, the results achieved superior values in training stage, while got affected in testing stage to classify the corresponding voice signal. Therefore, this research focused on two stages (i.e., feature selection and classification) which probably improvises the SER ability and helps to achieve better results in terms of accuracy, precision, recall and F1-score.

III. SER USING CP-LSTM WITH ATTENTION MECHANISM

The SER is a well-versed area which helps to perform communication among computers and humans. However, due to certain factors, overall efficiency of the recognition systems get diminished with poor accuracy. Therefore, more efforts have been put forward to enhance the efficiency during speech recognition but it had faced issues due to presence of undesired noises in the input signal. So, this research developed an effective framework using optimization-based feature selection and a novel deep learning approach. Initially, the data is

TABLE I. SUMMARY OF LITERATURE REVIEW

Author	Methodology	Disadvantage
Yildirim et al. [21]	Modified feature selection approach for speech emotion recognition.	Performing an exhaustive search for all subset space is infeasible.
Agarwal and Om [22]	An optimized Deep Neural Network (DNN) for speech emotion recognition.	Implementing a greater number of optimization techniques results in computational complexity.
Manohar and Logashanmugam [23]	A hybrid Deer Hunting with Adaptive Search (DH-AS) along with feature selection for recognizing the speech emotion.	The suggested approach was suitable only for the balanced dataset and complex due to its recurrent nature.
Mustaqeem and Kwon [24]	An optimal feature selection approach using two stream Deep Convolutional Neural Network (D-CNN).	The improper pre-processing leads to discrepancies and redundancies that diminishes the efficiency of the overall model.
Chattopadhyay et al. [25]	A hybrid feature selection approach using Clustering based Equilibrium Optimizer and Atom Search Optimization (CEOAS) for recognizing emotions from speech signals.	CEOAS exhibits premature convergence at the time of iterative searching process.
Kanwal et al. [26]	A Density based Spatial Clustering with Noise Genetic Algorithm (DGA) to recognize the type of emotion from the speech.	The PCA was computationally imprecise for huge datasets.
Barsainyan and Singh [27]	An optimized speech emotion recognition using 1D CNN.	The suggested framework was vulnerable to higher datasets due to limited model training.
Sun et al. [28]	A speech emotion recognition approach using Improved Masking based Empirical Model Decomposition and Convolutional Recurrent Neural Network (IMEMD-CRNN).	The mode mixing occurs which diminishes the capability of IMEMD to decompose the signal.
Otoni et al. [29]	A deep learning approach based on CNN and Long Short-Term Memory (LSTM) which is used to recognize the speech emotions based on meta learning approach.	The lack of feature selection leads to computational complexity and diminishes the efficiency of classification.
Nhat Truong Pham et al. [30]	A Hybrid Data Augmentation (HDA) with Modified Attention-Based Dilated Convolutional and Recurrent Neural Network (mADCRNN) methods are proposed for speech emotion recognition.	The number of layers and parameters in the suggested mADCRNN model made it complicated.
Zengzhao Chen et al. [31]	A parallel network for multi-scale SER based on connection attention mechanism (AMSNet) is proposed for speech emotion recognition.	Because of their less conspicuous characteristics, neutral emotions are less often recognized, since the characteristics of neutral emotions were not sufficiently evident and it was challenging to categorize the corresponding voice signal.
Mustaqeem Khan et al. [32]	Multimodal Speech Emotion Recognition (MSER).	In order to address the restriction of the complicated interaction between speech signals and transcripts, these investigations disregarded the limitations of past knowledge.
Dal Ri et al. [33]	An extensive validation using CNN for speech emotion recognition.	An appropriate selection of features was mandatory for further enhancing the recognition.
Y. Du et al. [34]	Heuristic Multimodal Real-Time Emotion Recognition (HMR-TER) approach.	A few tests with a limited range of accuracy were made available.
S. Debnath et al. [35]	An Audio-Visual Automatic Speech Recognition (AV-ASR) system was proposed to enhance the educational experience of those with physical disabilities by enabling hands-free computing.	As cluster size grows, accuracy decreases. This is due to the fact that a larger cluster size exhibits a dispersed representation of the data, lowering accuracy.
L. Pipiras et al. [36]	An Automatic Speech Recognition (ASR) system was developed for the Lithuanian language.	With shorter input sequences, the proposed model works merely well; however, it struggles with larger sequences.
A. Lauraitis et al. [37]	Bidirectional Long Short-Term Memory (BiLSTM) neural network and Wavelet Scattering Transform with Support Vector Machine (WST-SVM) classifier.	Although there is a significant association among phoneme and grapheme sequence lengths in the corpus of data, not all patterns have the same length.

acquisitioned from five different datasets and it is pre-processed to neglect the undesired information from the audio signal. After this, the pre-processed data is used to acquire the features. The extracted features are provided to phase of selecting features that takes place using Improved Jellyfish Optimization Algorithm (IJOA). As the final stage, classification takes place using the Convolutional Peephole Long Short-Term Memory (CP-LSTM) with attention mechanism. The overall process involved in the proposed Speech Emotion Recognition (SER) framework is presented in Fig. 1.

A. Dataset Description

The SER proposed in this research utilized five different types of datasets such as Berlin Database of Emotional Speech (EMO-DB) [38], Ryerson Audio Visual Database of Emotional Speech and Song

(RAVDESS) [39], Interactive Emotional Dyadic Motion Capture (IEMOCAP) [40], Surrey Audio-Visual Expressed Emotion (SAVEE) [41] and Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [42]. The aforementioned datasets are described as follows:

EMO-DB: It is obtained from Berlin emotion speech corpus that is recorded by a total of 10 actors where five were males and the remaining five were females. The dataset is comprised with a total of 535 audio files with an average time of 3-5 seconds and a sampling rate of 16kHz. Moreover, it is one of the vastly used datasets used in machine learning and deep learning techniques due to its clarity, which facilitates high recognition rates.

RAVDESS: It is one of the newly launched databases that is vastly utilized in evaluating the emotion of speech. This dataset is comprised

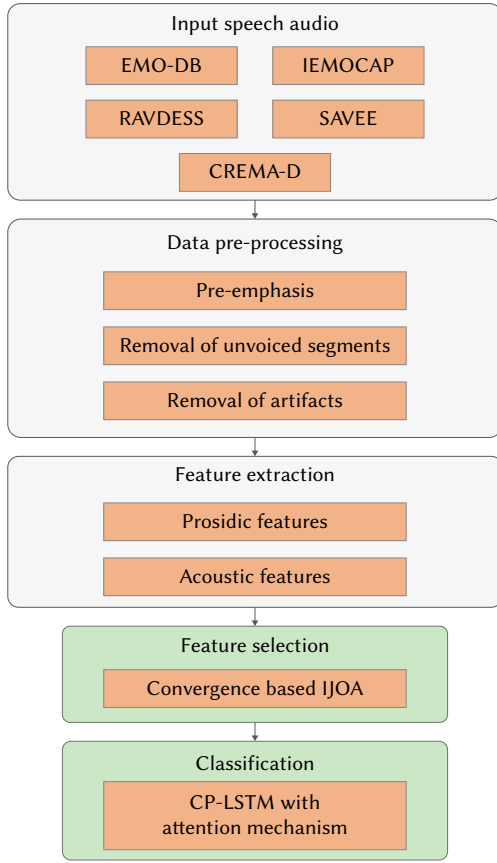


Fig. 1. Workflow of the proposed SER framework.

with total of eight kind of emotions which are recorded by 24 actors in a total of 12 sessions. The sampling rate of RAVDESS dataset is 48kHz with an average time of 3.5 s.

IEMOCAP: It is another type of vastly utilized SER datasets, which consists of improvised and scripted dialogues obtained from 10 actors in 5 sessions. This dataset is comprised with the audio clip of 12 hours and the utterances are annotated with emotion labels. IEMOCAP dataset is comprised with four stages such as anger, sad, neutral and happiness.

SAVEE: This dataset is comprised with a total of 480 utterances with varying emotions recorded by 4 actors at Centre for Vision, Speech and Signal Processing (CVSSP). Every speaker speaks 120 phonetically balanced English sentences based on 7 emotional classes.

CREMA-D: This dataset comprises 7,442 original clips from 91 actors. These clips have been gathered from 48 male and 43 female actors among the ages of 20 and 74 from a diversity of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified).

B. Data Pre-Processing

Next to data acquisition, the pre-processing is performed to neglect the undesired information from the raw data. In this research, the data pre-processing is performed using pre-emphasis, removal of artifacts and removal of unvoiced segments. The process involved in the aforementioned techniques are presented in the following sub-sections.

1. Pre-Emphasis

The speech signal pre-emphasis is the initial level of pre-processing in higher frequency. Here, the speech signal is transferred via high pass filter for enhancing the high-frequency band's amplitude. In specific, this research makes use of Finite Response (FIR) filter that

helps in the process of flattening high frequency audio signal. The process involved in pre-emphasis is represented in (1).

$$P(y) = 1 - hy^{-1} \quad (1)$$

Where the pre-emphasized output is represented as $P(y)$, the audio signal is represented as y and the filter co-efficient is represented as h .

2. Removal of Artifacts

After flattening the high frequency signals, the process of removing artifacts takes place which is used to remove the useless data and helps in effective recognition of emotions. In this research, the artifact removal takes place using the Fast Fourier Transform (FFT) that helps to remove the useless artifacts so the power spectrum of individual signals is evaluated using (2) as follows:

$$T_m(e) = \frac{1}{V} |S_m(e)|^2 \quad (2)$$

Where the spectral power of improved speech signals are represented as $S_m(e)$ and the total number of audio samples is represented as V .

3. Removal of Unvoiced Segments

Next to the stage of artifact removal, the unvoiced segments in the audio signals are removed, which increases the computational complexity while processing the audio signals. This research utilized Zero Cross Rate (ZCR) to remove the unvoiced segments. ZCR offers transition of signal over zero line which denotes noiseless measure in speech signals. ZCR is evaluated using (3) as follows:

$$Z_t = \frac{1}{2} \sum_{n=1}^N |sign(x[n]) - sign(x[n-1])| \quad (3)$$

Where $sign$ is the function that presents 1 for positive arguments and 0 for negative arguments. The time domain signal at the frame t is denoted as $x[n]$ which provides a measure of noiseless signal.

Thus, the pre-processing performed based on the fore mentioned techniques helps to remove the undesired information from the audio signal and helps to achieve better classification results thereby minimizing the complexity of the data. The pre-processed data is provided to the phase of extracting features.

C. Feature Extraction

Next to pre-processing, the feature extraction is performed to extract the relevant features from the pre-processed output. This extraction of features will be based on prosodic and acoustic features. The prosodic features such as energy, entropy, pitch and formants are utilized to extract the relevant features. In a similar way, the acoustic features such as Linear Predictive Coding (LPC), Linear Prediction Cepstral Co-efficient (LPCC), Mel-Frequency Cepstral co-efficient (MFCC), spectral flux and Zero Cross Rate (ZCR) are extracted from the pre-processed output.

1. Prosodic Features

The prosodic features such as energy, entropy and pitch are considered. Among the three, energy acts as the fundamental speech signal processing. In emotion recognition, energy plays a major role in recognizing the speech signals. Secondly, pitch acts as the periodic standard where high frequency harmonics are captured. The characteristics can be retained for every individual frame and it is pre-processed using short term analysis approach. The energy (E), pitch (P) and entropy (H) are evaluated based on (4)-(6) respectively.

$$E = \frac{1}{N} \sum_{n=m-N_w+1}^m |f_s(n; m)|^2 \quad (4)$$

$$P = \frac{F}{N} \arg\max_{\eta} \{ |r(\eta, m)| \}_{\eta=N_w(F_1/F_5)}^{\eta=N_w(F_h/F_5)} \quad (5)$$

$$H = - \sum_{i=1}^N s(i) \times \log [s(i)] \quad (6)$$

Where energy of speech frame is denoted as E , total count of frames is denoted as N and number of samples in the frame is denoted as N_w . The entropy is denoted as H , pitch frame is represented as P and the sampling frequency is represented as F_s . The low pitch frequency and high pitch frequency is represented as F_l and F_h respectively, measurement metric of glottal velocity is denoted as argmax .

2. Acoustic Features

Other than prosodic features, the acoustic features are considered while extracting the features. The fore mentioned acoustic features such as LPC, LPCC, MFCC, spectral flux, and ZCR are considered while extracting the acoustic based features.

LPC: The LPC computes the intensity of the input signal thereby eliminating the frequency of left over buzz. The remaining portion of the signal is achieved as residual signal that is represented in (7):

$$s(n) = \alpha_1 s(n-1) + \alpha_2 s(n-2) + \dots + \alpha_p s(n-p) \quad (7)$$

Where the linear co-efficient is denoted as $\alpha_1, \alpha_2, \dots, \alpha_p$.

LPCC: It is similar to the feature extraction performed by LPC where the Cepstral co-efficient is extracted from the features of LPC. Next to this, the representation of co-efficient is performed using the derivative of Fourier transform.

MFCC: It is the illustration of short time power spectrum of sound where the square magnitude of windowed speech signal is evaluated. The Mel scale filters along with log power spectrum is employed by overlapping the critical band filters. The co-efficient is assessed by performing Discrete Cosine Transform (DCT) of Mel-bin log energies is denoted in equation (8).

$$C_i(t) = \sum_{a=1}^B \log m_a(t) \cdot \cos\left(\frac{i(a-0.5)\pi}{b}\right) \quad (8)$$

Where, Mel cepstral coefficients are denoted as C_p , an amount of Mel filters at the filter bank is specified as B , $i = 1, 2, \dots, p$, amount of DCT points is denoted as p and the triangular filter bank function is represented as b .

Spectral flux: The speech signals which are pre-processed are obtained using the spectral flux to extract the various features of emotions. The spectral flux of the speech signal is denoted as SF_m which is evaluated using (9) as follows:

$$SF_m = \sum_{G=0}^{pt-1} (|E_m(G)|^2 - |E_{m-1}(G)|^2) \quad (9)$$

Where the spectrum value of the speech signal at frame m and $m-1$ in frequency bin G are represented as $E_m(G)$ and $E_{m-1}(G)$ respectively, and total amount of points in the spectrum is denoted as pt .

ZCR: It is a general type of feature insert which quantifies the amplitude of the speech signal, which has a zero value threshold in a particular time frame. ZCR has the ability to distinguish among the voiced and unvoiced signals and mathematically ZCR is evaluated based on (10) as follows:

$$ZCR = \frac{1}{T-1} \sum_{t=1}^{T-1} 1_{R<0} \quad (10)$$

Where the length of signal S is represented as T and $1_{R<0}$ is the indicator function.

Finally, the features which are extracted based on the prosodic features and the acoustic features are concatenated which is fed into the stage of feature selection to select the optimal features.

D. Feature Selection

Next to the stage of extraction of features, the selection of optimal features is performed with the help of proposed Improved Jellyfish Optimization Algorithm (IJOA). The existing Jellyfish Optimization Algorithm (JOA) is based on the searching pattern while searching for

the food. Initially, the jelly fish goes through ocean current then moves within the group based on two motions specified as type A and type B. A brief explanation about the iterative process involved in IJOA along with the enhancement made to select the relevant features are represented in the following sub-sections.

1. Overview of JOA

The movement of jelly fish is based on their active and passive behavior represented as type A motion and type B motion. The time control principle plays a major role in determining those two time varying motion types and this plays an important role in the development of controlling variation among type A and type B.

a) Initialization of Population

In general, the population of optimization algorithms are initialized in a random manner. This randomized initialization results in minimal precision and limited running value. So in JOA, the initialization is performed using logistic maps and randomness is generated using chaotic maps. This is represented in (11) as follows:

$$P_{i+1} = \eta P_i (1 - P_i) \quad (11)$$

Where logistic chaos value based on candidate's position is represented as P_i and the population at initial stage is represented as P_0 .

b) Behavior of Following Ocean Current

The direction of each variable for candidate solution from the position to optimal position is represented as current direction ($\overrightarrow{Direction}$) which is expressed in (12) as follows:

$$\overrightarrow{Direction} = \frac{1}{N} \sum \overrightarrow{Direction}_i = \frac{1}{N} \sum (P^* - e_c P_i) = P^* - e_c \frac{\sum P_i}{N} = P^* - e_c \mu = P^* - df \quad (12)$$

Where N , e_c and μ represents population of individual candidates, attraction factor and position of jelly fish correspondingly. The optimal position of individual candidate is represented as P^* and the variation among the optimal and the average location is represented as $df = e_c \mu$.

c) Movement of Jelly Fish

The movement of jelly fish is based on two types such as type A and type B. In type A, most of the candidate solutions do not show ability and in type B, the jelly fish starts the move in passage of time.

(1) Movement of type A

It is the type of passive movement where individual candidate changes along the own position which is updated using (13) as follows:

$$P_i(t+1) = P_i(t) + \gamma \times r_3 \times (U_b - L_b) \quad (13)$$

Where upper and lower limit of the search space is denoted as U_b and L_b correspondingly, movement factor is represented as γ and the value of r_3 lies among the range of (0,1).

(2) Movement of type B

It is the type of active movement where the individual candidate (j) is selected in a random manner. When the total quantity of food exceeds the location of selected candidates, the position of P_j exceeds the own location P_i . Every individual candidate migrates from one direction to another in search of food where the position of the candidate gets updated on the basis of (14)

$$P_i(t+1) = P_i(t) + \overrightarrow{step} \quad (14)$$

Where \overrightarrow{step} is calculated as expressed in equations (15) and (16).

$$\overrightarrow{step} = \text{rand}(0,1) \times \overrightarrow{DDirection} \quad (15)$$

$$\overrightarrow{DDirection} = \begin{cases} P_j(t) - P_i(t) & \text{if } f(P_i) \geq f(P_j) \\ P_i(t) - P_j(t) & \text{if } f(P_i) < f(P_j) \end{cases} \quad (16)$$

d) Time Control Mechanism

After the stage of capturing the movement, time control theory is utilized to adjust the varying candidate solutions. The candidate individuals present in the population of ocean current are denoted in (17).

$$C(t) = |(1 - t/T) \times (2 \times \text{rand}(0,1) - 1)| \quad (17)$$

Where the iterations at current stage and the maximum stage is denoted as t and T correspondingly.

2. Convergence Based IJOA

The existing JOA faced issues related to poor precision value and tends to fall into the local optima. So, this research introduced improvisation in JOA by introducing sine and cosine learning factors, and a premature convergence strategy. The sine and cosine learning is based on stage of exploration and the premature convergence strategy is based on exploitation stage. The inducement of learning strategy and the premature convergence strategy helps to overcome the issues related to poor precision and poor convergence rate.

a) Factors Based on Sine and Cosine Learning for Exploration

In the exploration phase, jelly fish performs type B motion in the whole population. The jelly fish learns from the other individual in current population that results in certain inappropriateness and lacks a proper exchange among the population. This results in placement of proper candidate solution and slows down the convergence speed. Therefore, the learning factors such as ω_1 and ω_2 based on sine and cosine functions are introduced to enhance the capability of jelly fish to learn from random and best individuals. The presentation of learning strategies in the exploration stage helps to enhance the quality of candidate solution by identifying the optimal location and helps to improvise the convergence speed. The mathematical equations of sine and cosine learning factor are represented as ω_1 and ω_2 which are presented in (18) and (19) as follows:

$$\omega_1 = 2 \cdot \sin \left[\left(1 - \frac{t}{T} \right) \cdot \pi / 2 \right] \quad (18)$$

$$\omega_2 = 2 \cdot \cos \left[\left(1 - \frac{t}{T} \right) \cdot \pi / 2 \right] \quad (19)$$

During the stage of updating the type B movement, the location of the jelly fish got updated as is denoted in (20)

$$P_i(t+1) = \omega_1 \cdot (P_i(t) + \overrightarrow{\text{step}}) + \omega_2 \cdot (P^* - P_i(t)) \quad (20)$$

Where the value of $\overrightarrow{\text{step}}$ is represented in (15) and (16). The actual JOA utilizes a random learning technique to learn from their individual. Moreover, the poor fitness functions of the jelly fish limit the speed of convergence. Then, the sine and cosine factors help to learn from random solutions and help to enhance the quality of solution with quick convergence rate.

b) Premature Convergence Strategy for Exploitation

The existing JOA experience low exploitation due to the inability of algorithm to accelerate the convergence for an optimal solution. Additionally, the capacity of swarm exploitation for accomplishing a search in the population is restricted. This problem occurs in exploitation, that is important in updating solutions and enhancing the search in the population. The capacity of the algorithm is restricted, when the small solution is enhanced, and the capacity of swarm for searching other area maximizes when the r is high. The mathematical expression for premature convergence rate is denoted in (21).

$$\vec{X}_i(t+1) = \vec{X}_i(t) + r \times (\vec{X}_{r1}(t) - \vec{X}_{r2}(t)) + (1-r) \times (X^* - \vec{X}_{r3}(t)) \quad (21)$$

Where the indices of three solutions which are picked up in a random manner from $r1$, $r2$ and $r3$, where the control parameter is represented as r . The value of control parameters lies in the range of 0 and 1 which is utilized in controlling the movement of current solution. The premature convergence strategy is used to accelerate the rate of convergence with randomly selected solutions from the population. This entire process involved in selecting the optimal features is performed using the proposed IJOA and emotion categorization from speech is carried out using CP-LSTM with attention mechanism.

E. Classification Using CP-LSTM With Attention Mechanism

The various forms of Long Short-Term Memory (LSTM) with various architectures are vastly utilized in the applications related to speech emotion recognition. The traditional architecture of LSTM is widely utilized in SER. However, there are issues because dependencies among the cells are not strong, which affects the overall classification efficiency. Then, this research introduced an effective classification approach using the proposed CP-LSTM with attention mechanism. Even with the closed output gate, the suggested CP-LSTM permits access to the previous cell state. Moreover, the content of previous memory cell helps to capture the complete dependency to improve the model accuracy. Also, the attention strategy is included in the final layer of CP-LSTM model that assists to choose the important term and capturing the complete pattern of training data. The addition of attention layer supports the model to generate syntactically and semantically [43]; the brief details about the traditional architecture of LSTM and the proposed CP-LSTM with attention mechanism for SER is explained in following sections.

1. Convolutional Peephole LSTM With Attention Mechanism

The LSTM is a kind of Recurrent Neural Network (RNN) which has the capability to hold and remember the information for a particular period of time. The LSTM is highly recommended in processing the sequences from the selected features. The architectural diagram of the LSTM model is presented in Fig. 2.

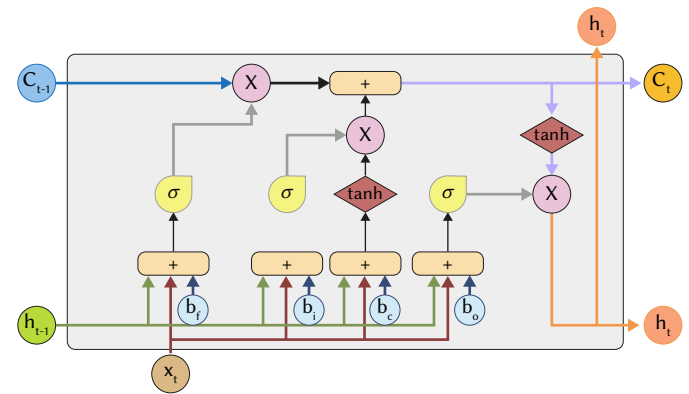


Fig. 2. Architectural diagram of the traditional LSTM model.

The architecture of LSTM is comprised with memory cells and three gates like forget, input and output gates in which data is stored using the memory cell and the control cell states are controlled using the remaining three gates. From Fig. 1, the line which is passed over architecture is represented as a memory pipe. The memory in previous state of memory line is denoted as C_{t-1} and the memory pipe is organized based on three gates. The sigmoid function is used to convert output of forget gate from 0 to 1. The bitwise summation integrates the temporary memory which is created using input gate with prior memory state along with the final memory which is represented as C_t .

The forget gate is a single layered neural network which takes place in different operation and regulate the memory content of previous cell. The forget gate obtains output state h_{t-1} , input vector x_t and bias input b_f . The final output generated from the forget gate is represented as f_t among the range of 0 to 1 and it is multiplied with C_{t-1} . The lower value of f_t prohibits the content of previous memory where the high value of f_t permit prior memory state to contribute the current state. The forget gate is represented in (22).

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (22)$$

The temporary memory state is comprised with sigmoid function and hyperbolic tangent function where sigmoid function produces output i_t , which is integrated in previous state, and the \tanh is a kind of activation function whose value lies among 0 and 1. The temporary memory with large \tanh function contributes better to the memory cell which is represented in (23) and (24).

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (23)$$

$$C_t = f_t C_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (24)$$

The output gate distinguishes the content h_t at a time t and the value of o_t in the range 0 to 1. When the value of o_t is equal to 1 then $h_t = C_t$ where the entire C_t is passed to next state h_t . The outcome through the output gate and hidden state of succeeding layer is represented as o_t and h_t which is represented in (25) and (26) respectively.

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (25)$$

$$h_t = o_t \tanh(C_t) \quad (26)$$

Based on the architecture of LSTM, the control gates are not incorporated with memory cell. Moreover, the output gate of LSTM is in closed state during training time, which may result in inappropriate classification result in SER. Therefore, this research introduced an advanced LSTM architecture known as CP-LSTM with attention mechanism to rectify the issues aroused while performing emotion recognition from speech.

a) CP-LSTM With Attention Mechanism for SER

The traditional LSTM architecture faced issues related to poor configuration among gates which prohibits the memory usage of prior memory states. The fore mentioned issue in traditional LSTM can be overwhelmed by introducing connection among individual gate and memory content which is referred as CP-LSTM. The peephole connection of CP-LSTM helps all the gates to access memory content. In CP-LSTM, the previous cell state C_{t-1} is linked with the controlling gate which is referred as peephole connections and the presence of this peephole permits an additional parameter and a memory state as the input of CP-LSTM. The inclusion of this additional input in each gate allows admission to memory content of prior cell state. The architectural diagram of CP-LSTM with attention mechanism is represented in Fig. 3. The architecture of CP-LSTM is similar to the architecture of traditional LSTM which is based on the mathematical expressions listed in (27)-(31)

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}C_{t-1} + b_i) \quad (27)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}C_{t-1} + b_f) \quad (28)$$

$$C_t = f_t C_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (29)$$

$$O_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}C_t + b_o) \quad (30)$$

$$h_t = o_t \tanh(C_t) \quad (31)$$

The proposed CP-LSTM utilized memory content of prior cell as input and the connection of C_{t-1} in CP-LSTM enhances the accuracy of prediction tasks. The attention mechanism present in the CP-LSTM architecture evaluates weight for every individual word which is based

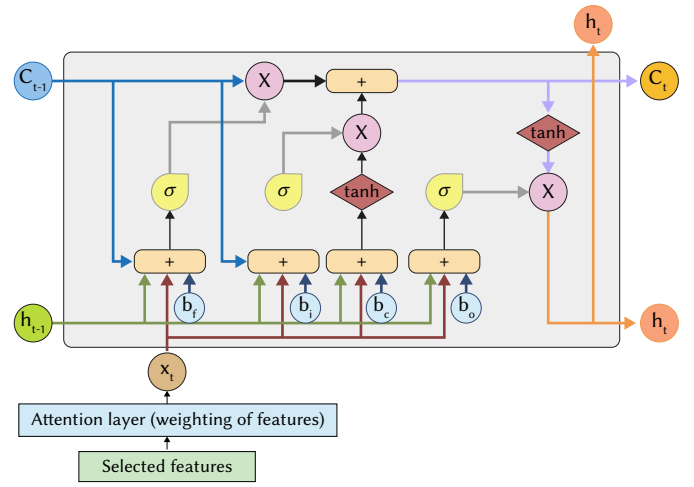


Fig. 3. Architectural diagram of CP-LSTM with attention layer.

on probability function that is used to evaluate the significant factors of input. The attention value of the signal is based on how much attention need to be paid while generating the recognition output from speech signals. When the speech signal is passed to the final hidden layer, the average weight of the speech signal is evaluated. Next to this, it is passed to the softmax layer with memory content of final hidden layer to predict emotion from speech. The attention weight of input features of speech signal is evaluated based on (32) as follows:

$$a_{y'_t}(t) = \frac{\exp(h_{x_t}^T h_{y'_t})}{\sum_t \exp(h_{x_t}^T h_{y'_t})} \quad (32)$$

Where last hidden layer which is created next to processing features is represented as $\exp(h_{x_t}^T)$. The attention mechanism in the CP-LSTM architecture helps to handle the problems of mapping large source to a static length. Moreover, the softmax layer is utilized to remove the outliers from output and map the vector. Thus, combination of proposed CP-LSTM with attention mechanism helps in an effective SER by recognizing the signal pattern, which results in better classification accuracy.

IV. RESULTS AND ANALYSIS

This section presents the analysis of the results achieved by evaluating the proposed CP-LSTM with attention mechanism with the state of art techniques and the existing approaches. Moreover, the efficiency of IJOA is evaluated with state of art optimization techniques while selecting the relevant features.

A. Experimental Setup and Evaluation Metrics

The efficiency of the CP-LSTM with attention mechanism based on the features selected using IJOA is implemented in python software and the system has configuration such as 16 GB RAM, i7 processor and windows 11 OS. The efficiency of the proposed approach is evaluated by considering the performance metrics such as accuracy, precision, recall and F1 score. Table II presents the performance metrics and the respective formula used while evaluation.

TABLE II. EVALUATION METRICS

Metrics	Formulae
Accuracy (A)	$A = \frac{TP + TN}{TP + TN + FP + FN}$
Precision (P)	$P = \frac{TP}{TP + FP}$
Recall (R)	$R = \frac{TP}{TP + FN}$
F-1 score (F)	$F = 2 \times \frac{P \times R}{P + R}$

Where TP and TN denote true positives and true negatives whereas the false positives and false negatives are represented as FP and FN respectively.

B. Performance Analysis

In this section, the performance of the proposed CP-LSTM with attention mechanism along with the proposed feature selection approach using IJOA is evaluated. The datasets EMO-DB, IEMOCAP, RAVDESS and SAVEE are used to evaluate the efficiency of IJOA and CP-LSTM. The quantitative analysis of the proposed method with all five datasets is presented in Table III.

TABLE III. QUANTITATIVE EVALUATION OF PROPOSED APPROACH

Datasets	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
EMO-DB	99.59	98.76	98.21	99.82
IEMOCAP	99.88	98.12	98.40	98.36
RAVDESS	99.54	99.83	99.57	98.36
SAVEE	98.89	98.57	98.19	98.43
CREMA-D	99.12	98.83	98.76	98.81

The Table III shows that the proposed approach accomplished an accuracy of 99.59%, 99.88%, 99.54%, 98.89% and 99.12% respectively for EMO-DB, IEMOCAP, RAVDESS, SAVEE and CREMA-D datasets. The following section presents the detailed analysis of the experimental outcome possessed by IJOA and CP-LSTM with attention mechanism for the different datasets.

1. Evaluation Based on Feature Selection

In this sub-section, the performance of IJOA utilized in feature selection is evaluated with the state of art optimization techniques such as Whale Optimization Algorithm (WOA), Grasshopper Optimization Algorithm (GOA) and Jellyfish Optimization Algorithm (JOA). The evaluation is performed by considering the five datasets utilized in this research. First, the efficiency of IJOA is evaluated with the existing optimization techniques such as WOA, GOA and JOA for EMO-DB dataset. The experimental outcome achieved while evaluating IJOA with existing ones for EMO-DB dataset is presented in table IV.

TABLE IV. EVALUATION OF IJOA FOR EMO-DB DATASET

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
WOA	97.35	96.22	96.81	96.05
GOA	97.62	96.75	96.87	96.27
JOA	96.81	97.20	96.15	96.69
IJOA	99.24	97.61	97.48	96.81

Experimental outcome from Table IV shows that IJOA achieved better results than the state of art optimization techniques. The accuracy of the proposed IJOA is 99.24% which is comparably higher than WOA, GOA and JOA with accuracies of 97.35%, 97.62% and 96.81%. Fig. 4 shows the graphical depiction of IJOA's performance for EMO-DB dataset.

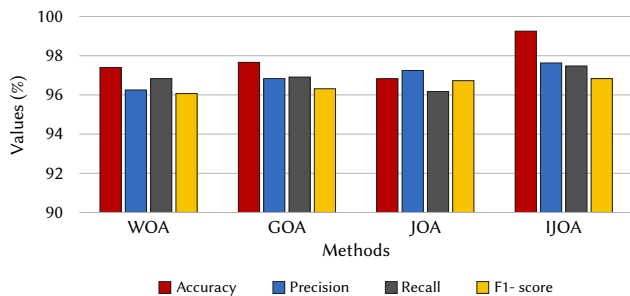


Fig. 4. Graphical representation of optimization techniques performance on EMO-DB dataset.

Secondly, the performance of IJOA is evaluated with WOA, GOA and JOA for IEMOCAP dataset. Table V shows the experimental results achieved while evaluating IJOA with state of art optimization techniques.

TABLE V. EVALUATION OF IJOA FOR IEMOCAP DATASET

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
WOA	96.12	95.23	95.39	95.28
GOA	95.66	96.21	96.58	96.59
JOA	97.18	96.68	96.19	96.17
IJOA	99.37	99.82	99.76	99.10

Table V shows IJOA achieved better results than state of art optimization techniques. The accuracy of the proposed IJOA is 99.37% which is comparably higher than WOA, GOA, and JOA with accuracies of 96.12%, 95.66% and 97.18% respectively. Fig. 5 shows the graphical depiction of IJOA's performance for IEMOCAP dataset.

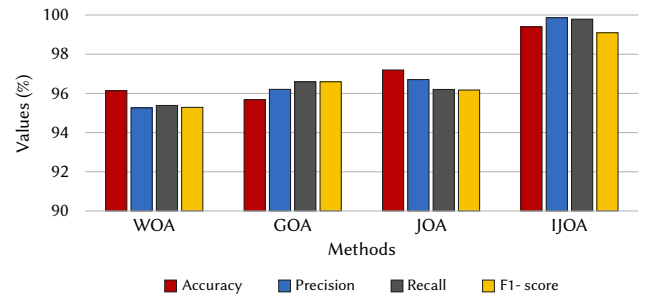


Fig. 5. Graphical representation of optimization techniques performance on IEMOCAP dataset.

Thirdly, the performance of IJOA is evaluated with WOA, GOA and JOA for RAVDESS dataset. Table VI shows the experimental results achieved while evaluating IJOA with state of art optimization techniques.

TABLE VI. EVALUATION OF IJOA FOR RAVDESS DATASET

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
WOA	94.21	96.90	96.01	96.22
GOA	96.59	97.55	95.89	95.97
JOA	97.82	96.81	96.27	96.80
IJOA	99.98	98.81	99.89	99.82

Table VI exhibits IJOA achieved better results than state of art optimization techniques. The accuracy of the proposed IJOA is 99.98% which is comparably higher than WOA, GOA, and JOA with accuracies of 94.21%, 96.59% and 97.82% respectively. Fig. 6 shows the graphical depiction of IJOA's performance for RAVDESS dataset.

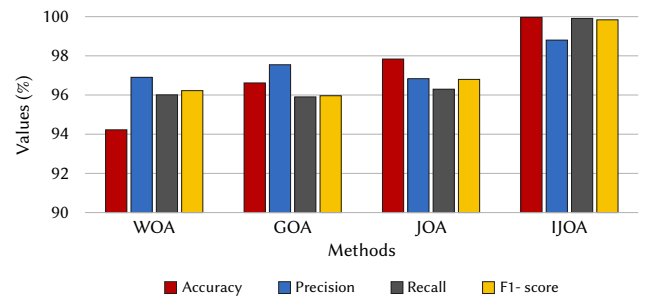


Fig. 6. Graphical representation of optimization techniques performance on RAVDESS dataset.

Finally, the performance of IJOA is evaluated with WOA, GOA and JOA for SAVEE dataset. Table VII shows the experimental outcome achieved while assessing IJOA with state of art optimization techniques.

TABLE VII. EVALUATION OF IJOA FOR SAVEE DATASET

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
WOA	96.82	96.78	96.82	96.29
GOA	97.11	97.80	97.99	97.87
JOA	97.56	97.25	97.69	97.58
IJOA	98.81	98.66	98.85	99.82

Table VII exhibits IJOA achieved better results than state of art optimization techniques. The accuracy of the proposed IJOA is 98.81% which is comparably higher than WOA, GOA, and JOA with accuracies of 96.82%, 97.11% and 97.56% respectively. Fig. 7 shows the graphical depiction of IJOA's performance for SAVEE dataset.

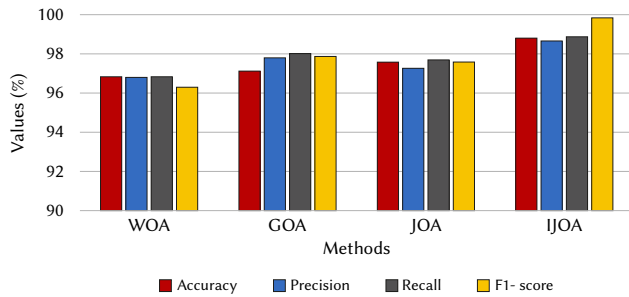


Fig. 7. Graphical representation of optimization techniques performance on SAVEE dataset.

The table VIII shows the experimental results accomplished while evaluating IJOA with state of art optimization methods on CREMA-D dataset.

TABLE VIII. EVALUATION OF IJOA FOR CREMA-D DATASET

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
WOA	95.38	94.18	95.44	95.26
GOA	95.91	95.73	96.19	96.43
JOA	96.56	96.49	96.38	96.71
IJOA	98.91	98.19	97.91	98.05

Table VIII shows that IJOA attained best results than state of art optimization methods. The accuracy of the proposed IJOA is 98.91% which is greater than WOA, GOA, and JOA with accuracies of 95.38%, 95.91% and 96.56%, respectively. Fig. 8 displays the graphical depiction of IJOA's performance for CREMA-D dataset.

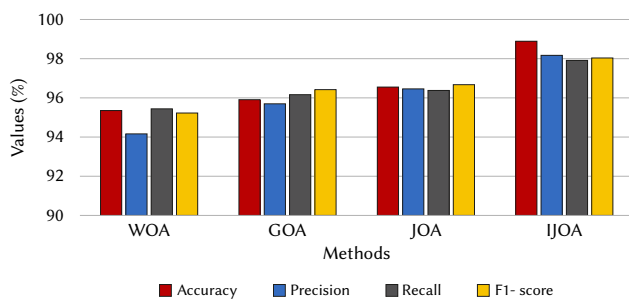


Fig. 8. Graphical representation of IJOA's performance on CREMA-D dataset.

The overall results from Table IV – VIII show that the proposed IJOA achieved better results in overall performance metrics when it

is compared with existing optimization techniques. For example, the accuracy of the proposed IJOA for SAVEE dataset is 98.81% which is comparably higher than WOA, GOA, and JOA methods. The optimal outcome of proposed IJOA is due to inclusion of the sine and cosine learning factors for exploration and the premature convergence strategy for stage of exploitation.

2. Evaluation Based on Classifier

In this sub-section, the performance of the proposed CP-LSTM with attention mechanism is evaluated based on its efficiency in SER with the five datasets utilized in this research. The performance of CP-LSTM with attention mechanism is evaluated with Recurrent Neural Network (RNN), Deep Belief Network (DBN) and Long Short Term Memory (LSTM) with attention mechanism. The performance is evaluated with the five publicly available datasets EMO-DB, IEMOCAP, RAVDESS, SAVEE and CREMA-D. Table IX shows the results achieved by the suggested classifier for EMO-DB dataset.

TABLE IX. COMPARISON OF DIFFERENT CLASSIFIERS WITH ATTENTION MECHANISM FOR EMO-DB DATASET

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
RNN	96.11	95.23	95.90	95.57
DBN	95.26	96.20	96.96	96.58
LSTM	97.02	96.53	96.62	96.57
CP-LSTM	99.59	98.76	98.21	99.82

The classification accuracy of CP-LSTM with attention mechanism is 99.59% which is comparably higher than the existing RNN, DBN and LSTM with accuracies of 96.11%, 95.26% and 97.02%.

Secondly, the performance of CP-LSTM with attention mechanism is evaluated for IEMOCAP dataset. Table X shows the results achieved by the suggested classifier for IEMOCAP dataset.

TABLE X. COMPARISON OF DIFFERENT CLASSIFIERS WITH ATTENTION MECHANISM FOR IEMOCAP DATASET

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
RNN	96.35	97.91	96.12	96.61
DBN	96.82	98.53	97.34	96.62
LSTM	97.25	97.75	97.96	97.42
CP-LSTM	99.88	98.12	98.40	98.36

The classification accuracy of CP-LSTM with attention mechanism is 99.88% which is comparably higher than the existing RNN, DBN and LSTM with accuracies of 96.35%, 97.91% and 97.25%.

Thirdly, the performance of CP-LSTM with attention mechanism is evaluated for RAVDESS dataset. Table XI shows the results achieved by the suggested classifier for RAVDESS dataset.

TABLE XI. COMPARISON OF DIFFERENT CLASSIFIERS WITH ATTENTION MECHANISM FOR RAVDESS DATASET

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
RNN	97.23	97.87	98.33	97.59
DBN	98.28	95.64	95.80	95.87
LSTM	97.53	98.82	98.97	98.48
CP-LSTM	99.54	99.83	99.57	98.36

The classification accuracy of CP-LSTM with attention mechanism is 99.54% which is comparably higher than the existing RNN, DBN and LSTM with accuracies of 97.23%, 98.28% and 97.53%. Table XII shows the results achieved by the suggested classifier for SAVEE dataset.

TABLE XII. COMPARISON OF DIFFERENT CLASSIFIERS WITH ATTENTION MECHANISM FOR SAVEE DATASET

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
RNN	96.21	95.09	97.56	96.41
DBN	96.09	97.02	96.78	96.89
LSTM	97.78	97.54	96.99	97.57
CP-LSTM	98.89	98.57	98.19	98.43

The classification accuracy of CP-LSTM with attention mechanism is 98.89% which is comparably higher than the existing RNN, DBN and LSTM with accuracies of 96.21%, 96.09% and 97.78%. Finally, the performance of CP-LSTM with attention mechanism is evaluated for CREMA-D dataset which is shown in Table XIII.

TABLE XIII. COMPARISON OF DIFFERENT CLASSIFIERS WITH ATTENTION MECHANISM FOR CREMA-D DATASET

Method	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
RNN	95.83	95.16	96.16	95.93
DBN	96.76	96.43	96.82	96.59
LSTM	97.98	97.59	97.38	97.17
CP-LSTM	99.12	98.83	98.76	98.81

The classification accuracy of CP-LSTM with attention mechanism is 99.12% which is comparably higher than the existing RNN, DBN and LSTM with accuracies of 95.83%, 96.76% and 97.98%.

Table IX - Table XIII show the results achieved while evaluating the proposed classifier with the five publicly available dataset utilized in this research. The performance of the proposed classifier seems to be high when it is compared with the existing state of art classifiers. The outstanding result of the proposed classification approach is due the peephole connection of CP-LSTM helps all the gates to evaluate memory content. In CP-LSTM, the prior cell state linked with the controlling gate represented as peephole connections and the presence of this peephole permits an additional parameter and a memory state as the input of CP-LSTM. The attention mechanism in the CP-LSTM architecture helps to handle problems of mapping large fixed length. Thus, combination of the proposed CP-LSTM with attention mechanism helps in an effective SER by recognizing the signal pattern and results in better classification accuracy.

C. Comparative Analysis

In this section, the efficiency of the proposed classifier is assessed with existing techniques based on SER. The comparison is performed with existing techniques such as DNN-DHO [22], DH-AS [23], D-CNN [24], CEOAS [25], CNN-LSTM [29] and CNN [33]. Table XIV shows the results achieved while evaluating the suggested approach with existing techniques and different datasets.

In an overall, the suggested method attained better results than existing ones due to the effective feature selection performed using IJOA and the efficient classification performed using the proposed CP-LSTM with attention mechanism. For example, the classification accuracy of CP-LSTM for RAVDESS dataset is 99.54% which is comparably higher than DNN-DHO, DH-AS, D-CNN and CNN-LSTM with accuracies of 97.5%, 95.6%, 85% and 97.01% respectively.

D. Discussion

This sub-section provides a brief discussion about the results achieved while evaluating CP-LSTM along with its advantages. In performance analysis, the efficiency of CP-LSTM with attention mechanism is evaluated with five publicly available datasets (EMO-DB, IEMOCAP, RAVDESS, SAVEE and CREMA-D). The proposed IJOA used for feature selection outperforms state of art optimization techniques

TABLE XIV. COMPARATIVE TABLE

Method	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F-1 score (%)
DNN-DHO [22]	RAVDESS	97.5	97.52	97.75	97.19
	DH-AS [23]	95.73	98.21	94.6	96.37
D-CNN [24]	RAVDESS	95.6	98.33	94.26	96.25
	EMO-DB	95	95.5	85	85.2
CEOAS [25]	SAVEE	82	74.8	72.9	73.7
	RAVDESS	85	85.4	85.4	85.4
	EMO-DB	98.72	DNA	DNA	DNA
	SAVEE	98.01	DNA	DNA	DNA
CNN-LSTM [29]	IEMOCAP	74.25	DNA	DNA	DNA
	SAVEE	90.62	DNA	DNA	DNA
CNN [33]	RAVDESS	97.01	DNA	DNA	DNA
	IEMOCAP	63	DNA	DNA	DNA
CP-LSTM with attention mechanism	RAVDESS	83	DNA	DNA	DNA
	CREMA-D	68	DNA	DNA	DNA
	EMO-DB	99.59	98.76	98.21	99.82
	SAVEE	98.89	98.57	98.19	98.43
	IEMOCAP	99.88	98.12	98.40	98.36
	RAVDESS	99.54	99.83	99.57	98.36
	CREMA-D	99.12	98.83	98.76	98.81

*DNA- data not available

such as WOA, GOA, and JOA. Similarly, CP-LSTM outperforms state of art classification approaches such as RNN, DBN and LSTM. The comparative analysis demonstrates that the CP-LSTM with attention mechanism offers better performance than the existing techniques such as DNN-DHO, DH-AS, D-CNN, CEOAS and CNN-LSTM. The sine and cosine functions incorporated in exploration stage of IJOA help to improve the candidate solution's quality while searching for optimum subset of features. This enhanced searching capacity of IJOA helps to mitigate the irrelevant features which is helpful in enhancing the recognition. On the contrary, the designed CP-LSTM allows the access to the previous cell state that leads to acquire the complete dependency for enhancing the SER performances. Further, the attention strategy included in the final layer of CP-LSTM is used for selecting the significant term and obtaining the complete pattern of training information for an additional improvement in the SER.

V. CONCLUSION

This research study introduced an effective classification approach using CP-LSTM with attention mechanism where the selected features are provided as input using IJOA. The major contribution of this research is to produce a robust framework for speech emotion recognition using five datasets namely, EMO-DB, IEMOCAP, RAVDESS, SAVEE and CREMA-D. These five datasets are pre-processed using pre-emphasis, removal of artifacts and removal of unvoiced segments to remove the undesired information from the signals. Then, the extraction of features takes place using prosodic features and acoustic features. The prosodic features such as energy, entropy, pitch and formants are utilized to extract the relevant features. In a similar way, the acoustic features such as LPC, LPCC, MFCC, spectral flux and ZCR are extracted from pre-processed output. Next to the feature extraction, the IJOA is used to select the relevant features which are fed into the classification that is performed using CP-LSTM with attention mechanism. The experimental results exhibit effectiveness of suggested approach by analyzing the classification accuracy for EMO-DB, IEMOCAP, RAVDESS and SAVEE datasets, which are 99.59%, 99.88%, 99.54% and 98.89% respectively, which is

comparably higher than the existing state-of-the-art models. Similarly, while working on RAVDESS dataset, the existing DNN-DHO, DH-AS, D-CNN and CNN-LSTM attained an accuracy of 97.5%, 95.6%, 85% and 97.01% respectively which is lower than proposed approach which has 99.54% accuracy. In the future, the suggested framework will be further implemented for time scenarios in industries and healthcare to analyze the classification accuracy.

NOTATION LIST

symbol	Description
$P(y)$	Pre-emphasized output
y	Audio signal
h	Filter co-efficient
$S_m(e)$	Spectral power of improved speech signals
V	Total number of audio samples
$sign$	Function
$x[n]$	Time domain signal
E	Energy of speech frame
P	Pitch frame
H	Entropy
N	Total count of frames
F_s	Sampling frequency
F_l	Low pitch frequency
F_h	High pitch frequency
$\arg\max$	Measurement metric of glottal velocity
$S(n)$	Residual signal
$\alpha_1, \alpha_1, \dots, \alpha_p$	Linear co-efficient
b	Triangular filter bank function
SF_m	Spectral flux of the speech signal
$E_m(G)$	Spectrum value of speech signal
S	Signal
T	Length of signal
$1_{R<0}$	Indicator function
P_i	Candidate's position
P_0	Population at initial stage is represented as
$(Direction)$	Direction of each variable
N	Population of individual candidates
e_c	Attraction factor
μ	Position of jelly fish
P^*	Optimal position of individual candidate
df	Variation among the optimal and the average location
U_b and L_b	Upper and lower limit
γ	Movement factor
j	Individual candidate
P_j	Position
P_i	Location
t	Current iterations
T	Maximum iteration
ω_1 and ω_2	Sine and cosine learning factor
r_1, r_2 and r_3	Random variables
r	Control parameter
C_{t-1}	Previous cell state
$\exp(h_{x_t}^T)$	Last hidden layer
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

DECLARATIONS

Funding: This work has been supported by the project PCI2022-134990-2 (MARTINI) of the CHISTERA IV Cofund 2021 program; by MCIN/AEI/10.13039/501100011033/ and European Union NextGenerationEU/PRTR for XAI-Disinfodemics (PLEC 2021-007681) grant, by European Comission under IBERIFIER Plus - Iberian Digital Media Observatory (DIGITAL-2023-DEPLOY- 04-EDMO-HUBS 101158511); and by EMIF managed by the Calouste Gulbenkian Foundation, in the project MuseAI.

Data Availability: The datasets generated during and/or analysed during the current study are available in the [EMO-DB], [RAVDESS], [IEMOCAP] and [SAVEE] datasets.

- EMO-DB dataset: <https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb>
- RAVDESS dataset: <https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio>
- IEMOCAP dataset: <https://sail.usc.edu/iemocap/>
- SAVEE dataset: <http://kahlan.eps.surrey.ac.uk/savee/>
- CREMA-D dataset: <https://github.com/CheyneyComputerScience/CREMA-D>

REFERENCES

- [1] A. A. Abdelhamid, E. S. M. El-Kenawy, B. Alotaibi, G. M. Amer, M. Y. Abdelkader, A. Ibrahim, and M. M. Eid, "Robust speech emotion recognition using CNN+ LSTM based on stochastic fractal search optimization algorithm," *IEEE Access*, vol. 10, pp. 49265-49284, 2022, <https://doi.org/10.1109/ACCESS.2022.3172954>.
- [2] Z. Chen, J. Li, H. Liu, X. Wang, H. Wang, and Q. Zheng, "Learning multi-scale features for speech emotion recognition with connection attention mechanism," *Expert Systems with Applications*, vol. 214, p. 118943, 2023, <https://doi.org/10.1016/j.eswa.2022.118943>.
- [3] A. Aggarwal, A. Srivastava, A. Agarwal, N. Chahal, D. Singh, A. A. Alnuaim, and H. N. Lee, "Two-way feature extraction for speech emotion recognition using deep learning," *Sensors*, vol. 22, no. 6, p. 2378, 2022, <https://doi.org/10.3390/s22062378>.
- [4] S. Kumar, M. A. Haq, A. Jain, C. A. Jason, N. R. Moparthi, N. Mittal, and Z. S. Alzamil, "Multilayer Neural Network Based Speech Emotion Recognition for Smart Assistance," *Computers, Materials & Continua*, vol. 75, no. 1, 2023, <https://doi.org/10.32604/cmc.2023.028631>.
- [5] S. Jothimani and K. Premalatha, "MFF-SAUG: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network," *Chaos, Solitons Fractals*, vol. 162, p. 112512, 2022, <https://doi.org/10.1016/j.chaos.2022.112512>.
- [6] D. Yang, S. Huang, Y. Liu, and L. Zhang, "Contextual and cross-modal interaction for multi-modal speech emotion recognition," *IEEE signal processing letters*, vol. 29, pp. 2093-2097, 2022, <https://doi.org/10.1109/LSP.2022.3210836>.
- [7] J. Lei, X. Zhu, and Y. Wang, "BAT: Block and token self-attention for speech emotion recognition," *Neural Networks*, vol. 156, pp. 67-80, 2022, <https://doi.org/10.1016/j.neunet.2022.09.022>.
- [8] B. Maji and M. Swain, "Advanced fusion-based speech emotion recognition system using a dual-attention mechanism with conv-caps and bi-gru features," *Electronics*, vol. 11, no. 9 p. 1328, 2022, <https://doi.org/10.3390/electronics11091328>.
- [9] Y. Liu, H. Sun, W. Guan, Y. Xia, and Z. Zhao, Multi-modal speech emotion recognition using self-attention mechanism and multi-scale fusion framework," *Speech Communication*, vol. 139, pp. 1-9, 2022, <https://doi.org/10.1016/j.specom.2022.02.006>
- [10] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1912-1926, 2022, <https://doi.org/10.1109/TAFFC.2022.3167013>.
- [11] S. Kakuba, A. Poullose, and D. S. Han, "Attention-based multi-learning approach for speech emotion recognition with dilated convolution,"

- IEEE Access*, vol. 10, pp. 122302-122313, 2022, <https://doi.org/10.1109/ACCESS.2022.3223705>.
- [12] P. R. Prakash, D. Anuradha, J. Iqbal, M. G. Galety, R. Singh, and S. Neelakandan, "A novel convolutional neural network with gated recurrent unit for automated speech emotion recognition and classification," *Journal of Control and Decision*, vol. 10, no. 1 pp. 54-63, 2023, <https://doi.org/10.1080/23307706.2022.2085198>.
- [13] R. K. Das, N. Islam, M. R. Ahmed, S. Islam, S. Shatabda, and A. M. Islam, "BanglaSER: A speech emotion recognition dataset for the Bangla language," *Data in Brief*, vol. 42, p. 108091, 2022, <https://doi.org/10.1016/j.dib.2022.108091>.
- [14] B. B. Al-onazi, M. A. Nauman, R. Jahangir, M. M. Malik, E. H. Alkhamash, and A. M. Elshewey, "Transformer-based multilingual speech emotion recognition using data augmentation and feature fusion," *Applied Sciences*, vol. 12, no. 18 p. 9188, 2022, <https://doi.org/10.3390/app12189188>.
- [15] J. Li, X. Zhang, L. Huang, F. Li, S. Duan, and Y. Sun, "Speech Emotion Recognition Using a Dual-Channel Complementary Spectrogram and the CNN-SSAE Neutral Network," *Applied Sciences*, vol. 12, no. 19, p. 9518, 2022, <https://doi.org/10.3390/app12199518>.
- [16] J. Ancilin and A. Milton, "Improved speech emotion recognition with Mel frequency magnitude coefficient," *Applied Acoustics*, vol. 179, p. 108046, 2021, <https://doi.org/10.1016/j.apacoust.2021.108046>.
- [17] H. A. Abdulmohsin, "A new proposed statistical feature extraction method in speech emotion recognition," *Computers & Electrical Engineering*, vol. 93, p. 107172, 2021, <https://doi.org/10.1016/j.compeleceng.2021.107172>.
- [18] C. Zhang and L. Xue, "Autoencoder with emotion embedding for speech emotion recognition," *IEEE access*, vol. 9, pp. 51231-51241, 2021, <https://doi.org/10.1109/ACCESS.2021.3069818>.
- [19] P. T. Krishnan, A. N. Joseph Raj, and V. Rajangam, "Emotion classification from speech signal based on empirical mode decomposition and non-linear features: Speech emotion recognition," *Complex & Intelligent Systems*, vol. 7, pp. 1919-1934 2021, <https://doi.org/10.1007/s40747-021-00295-z>.
- [20] K. S. Chintalapudi, I. A. K. Patan, H. V. Sontineni, V. S. K. Muvvala, S. V. Gangashetty, and A. K. Dubey, "Speech Emotion Recognition Using Deep Learning," In *2023 International Conference on Computer Communication and Informatics (ICCCI)*, IEEE, Coimbatore, India, 2023, pp. 1-5, <https://doi.org/10.1109/ICCCI56745.2023.10128612>.
- [21] S. Yildirim, Y. Kaya, and F. Kılıç, "A modified feature selection method based on metaheuristic algorithms for speech emotion recognition," *Applied Acoustics*, vol. 173, p. 107721, 2021, <https://doi.org/10.1016/j.apacoust.2020.107721>.
- [22] G. Agarwal and H. Om, "Performance of deer hunting optimization based deep learning algorithm for speech emotion recognition," *Multimedia Tools and Applications*, vol. 80, pp. 9961-9992, 2021, <https://doi.org/10.1007/s11042-020-10118-x>.
- [23] K. Manohar and E. Logashanmugam, "Hybrid deep learning with optimal feature selection for speech emotion recognition using improved meta-heuristic algorithm," *Knowledge-Based Systems*, vol. 246, p. 108659, 2022, <https://doi.org/10.1016/j.knosys.2022.108659>.
- [24] Mustaqeem and S. Kwon, "Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network," *International Journal of Intelligent Systems*, vol. 36, no. 9 pp. 5116-5135, 2021, <https://doi.org/10.1002/int.22505>.
- [25] S. Chattopadhyay, A. Dey, P. K. Singh, A. Ahmadian, and R. Sarkar, "A feature selection model for speech emotion recognition using clustering-based population generation with hybrid of equilibrium optimizer and atom search optimization algorithm," *Multimedia Tools and Applications*, vol. 82, no. 7, pp. 9693-9726, 2023, <https://doi.org/10.1007/s11042-021-11839-3>.
- [26] S. Kanwal and S. Asghar, "Speech emotion recognition using clustering based GA-optimized feature set," *IEEE Access*, vol. 9, pp. 125830-125842, 2021, <https://doi.org/10.1109/ACCESS.2021.3111659>.
- [27] N. Barsainyan and D. K. Singh, "Optimized cross-corpus speech emotion recognition framework based on normalized 1D convolutional neural network with data augmentation and feature selection," *International Journal of Speech Technology*, vol. 26, no. 4, pp. 947-961, 2023, <https://doi.org/10.1007/s10772-023-10063-8>.
- [28] C. Sun, H. Li, and L. Ma, "Speech emotion recognition based on improved masking EMD and convolutional recurrent neural network," *Frontiers in Psychology*, vol. 13, p. 1075624, 2023, <https://doi.org/10.3389/fpsyg.2022.1075624>.
- [29] L. T. C. Ottoni, A. L. C. Ottoni, and J. D. J. F. Cerqueira, "A Deep Learning Approach for Speech Emotion Recognition Optimization Using Meta-Learning," *Electronics*, vol. 12, no. 23, p. 4859, 2023, <https://doi.org/10.3390/electronics12234859>.
- [30] N. T. Pham, D. N. M. Dang, N. D. Nguyen, T. T. Nguyen, H. Nguyen, B. Manavalan, V. P. Lim, and S. D. Nguyen, "Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition," *Expert Systems with Applications*, vol. 230, p. 120608, 2023.
- [31] Z. Chen, J. Li, H. Liu, X. Wang, H. Wang, and Q. Zheng, "Learning multi-scale features for speech emotion recognition with connection attention mechanism," *Expert Systems with Applications*, vol. 214, p.118943, 2023.
- [32] M. Khan, W. Gueaieb, A. El Saddik, and S. Kwon, "MSER: Multimodal speech emotion recognition using cross-attention with deep fusion," *Expert Systems with Applications*, vol. 245, p.122946, 2024.
- [33] F. A. Dal Ri, F. C. Ciardi, and N. Conci, "Speech Emotion Recognition and Deep Learning: An Extensive Validation using Convolutional Neural Networks," *IEEE Access*, vol. 11, pp. 116638-116649, 2023.
- [34] Y. Du, R. G. Crespo, and O. S. Martínez, "Human emotion recognition for enhanced performance evaluation in e-learning," *Progress in Artificial Intelligence*, vol. 12, no. 2, pp. 199-211, 2023.
- [35] S. Debnath, P. Roy, S. Namasudra, and R. G. Crespo, "Audio-Visual Automatic Speech Recognition Towards Education for Disabilities," *Journal of Autism and Developmental Disorders*, vol. 53, no. 9, pp. 3581-3594, 2023.
- [36] L. Pipiras, R. Maskeliūnas, and R. Damaševičius, "Lithuanian speech recognition using purely phonetic deep learning," *Computers*, vol. 8, no. 4, p. 76, 2019.
- [37] A. Lauraitis, R. Maskeliūnas, R. Damaševičius, and T. Krilavičius, "Detection of speech impairments using cepstrum, auditory spectrogram and wavelet time scattering domain features," *IEEE Access*, vol. 8, pp. 96162-96172, 2020.
- [38] Link for EMO-DB dataset: <https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb>
- [39] Link for RAVDESS dataset: <https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio>
- [40] Link for IEMOCAP dataset: <https://sail.usc.edu/iemocap/>
- [41] Link for SAVEE dataset: <http://kahlan.eps.surrey.ac.uk/savee/>
- [42] Link for CREMA-D dataset: <https://github.com/CheyneyComputerScience/CREMA-D>
- [43] M. M. Rahman and F. H. Siddiqui, "Multi-layered attentional peephole convolutional LSTM for abstractive text summarization," *Etri Journal*, vol. 43, no. 2, pp. 288-298, 2021.



Ramya Paramasivam

Ramya Paramasivam received her M. Tech. degree in Computer Science and Engineering and Ph.D. in Information and Communication Engineering from Anna University in the years 2005 and 2017 respectively. Her research areas are Artificial Intelligence and Wireless Communication. She is currently an Associate Professor in the Department of Computer Science Engineering, Mahendra Engineering College, for the last 6 Years. Dr.P.Ramya has over 19 years' experience in teaching, research and industry. She has taught many courses in Computer Science and Engineering at post graduate and undergraduate levels. Dr.P.Ramya has published and presented papers at many international journals/conferences / symposiums and delivered many invited / plenary / expert lectures in India.



K. Lavanya

K. Lavanya received the Ph.D. degree in Information and communication from Anna University, Chennai, India. She is currently an Assistant Professor in the department of Electronics and Communication Engineering, Velammal Engineering College, Chennai. She has a strong academic teaching and research experience of more than 18 years. Her research interests include Internet of Things, machine learning techniques, cryptography and network security.



Parameshachari Bidare Divakarachari

Parameshachari Bidare Divakarachari is currently working as a professor in the Department of Electronics and Communication Engineering at Nitte Meenakshi Institute of Technology, Bangalore, India; affiliated to Visvesvaraya Technological University (VTU), Belagavi, Karnataka, India. He has around 19 years of experience and has published over 200+ articles in SCI, SCOPUS, and other indexed journals and also in conferences. He serves as editorial board member, associate editor, academic editor, guest editor, and reviewer for various reputed indexed journals. He is also the founder chair for IEEE Information Theory Society, Bangalore chapter and IEEE Mysore Subsection. He is also the SAC Chair, IEEE Bangalore Section.



David Camacho

David Camacho (Senior Member, IEEE) received the Ph.D. degree (with Hons.) in computer science from Universidad Carlos III de Madrid, Getafe, Spain, in 2001. He is currently a Full Professor with Computer Systems Engineering Department, Universidad Politécnica de Madrid (UPM), Madrid, Spain, and the Head of the Applied Intelligence and Data Analysis Research Group, UPM. He has authored or coauthored more than 300 journals, books, and conference papers. His research interests include machine learning (clustering/deep learning), computational intelligence (evolutionary computation, swarm intelligence), social network analysis, fake news and disinformation analysis. He has participated/led more than 50 research projects (Spanish and European: H2020, DG Justice, ISFP, and Erasmus+), related to the design and application of artificial intelligence methods for data mining and optimization for problems emerging in industrial scenarios, aeronautics, aerospace engineering, cybercrime/cyber intelligence, social networks applications, or video games among others. He was the recipient of the best thesis award in Computer Science for his Ph.D. degree.

Optimal Target-Oriented Knowledge Transportation for Aspect-Based Multimodal Sentiment Analysis

Linhao Zhang^{1,2,3}, Li Jin^{1*}, Guangluan Xu¹, Xiaoyu Li¹, Xian Sun¹, Zequn Zhang¹, Yanan Zhang⁴, Qi Li⁵

¹ Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing (China)

² University of Chinese Academy of Sciences, Beijing (China)

³ School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing (China)

⁴ Sichuan University, Chengdu (China)

⁵ Faculty of Psychology, Beijing Normal University, Beijing (China)

* Corresponding author: jinlimails@gmail.com

Received 9 February 2023 | Accepted 6 November 2023 | Early Access 13 February 2024



ABSTRACT

Aspect-based multimodal sentiment analysis under social media scenario aims to identify the sentiment polarities of each aspect term, which are mentioned in a piece of multimodal user-generated content. Previous approaches for this interdisciplinary multimodal task mainly rely on coarse-grained fusion mechanisms from the data-level or decision-level, which have the following three shortcomings: (1) ignoring the category knowledge of the sentiment target (mentioned in the text) in visual information. (2) unable to assess the importance of maintaining target interaction during the unimodal encoding process, which results in indiscriminative representations considering various aspect terms. (3) suffering from the semantic gap between multiple modalities. To tackle the above challenging issues, we propose an optimal target-oriented knowledge transportation network (OtarNet) for this task. Firstly, the visual category knowledge is explicitly transported through input space translation and reformulation. Secondly, with the reformulated knowledge containing the target and category information, the target sensitivity is well maintained in the unimodal representations through a multistage target-oriented interaction mechanism. Finally, to eliminate the distributional modality gap by integrating complementary knowledge, the target-sensitive features of multiple modalities are implicitly transported based on the optimal transport interaction module. Our model achieves state-of-the-art performance on three benchmark datasets: Twitter-15, Twitter-17 and Yelp, together with the extensive ablation study demonstrating the superiority and effectiveness of OtarNet.

KEYWORDS

Aspect-Based
Multimodal Sentiment
Analysis, Optimal
Transport, Social Media
Opinion Mining.

DOI: 10.9781/ijimai.2024.02.005

I. INTRODUCTION

SOCIAL media websites provide interactive platforms to facilitate the creation and sharing of individuals' expressions through multiple social activities (for example, 'like', 'reply', 'retweet', '@', 'share' in Twitter) [1]. Fine-grained sentiment analysis over these user generated content (UGC) in social websites (e.g., Twitter, Flickr) are effective in understanding public opinions toward social hotspots or figures, and it has drawn increasing recent attention in both academia and industry [2]. For example, socialists and psychologists have strong interests in understanding individual reactions toward specific social issues. Companies are willing to acquire online evaluations of their products as feedback to make further improvements. Therefore, how to incorporate heterogeneous multimodal information to conduct fine-grained sentiment analysis over the mentioned aspect terms has become an emerging interdisciplinary research problem, proposed as

Aspect-Based Multimodal Sentiment Analysis (ABMSA) [3]-[5].

Despite the well-established research fields of multimodal learning and affective computing, there are under-researched challenges for the aspect-based multimodal sentiment analysis (ABMSA) toward social media user-generated content (UGC): (1) Due to the viral nature of internet posts, sentences in social media UGC are always shorter, more informative and informal compared to the well-organized reviews used for traditional affective computing. (2) The visual information is much noisier with multiple objects for UGC than for videos of human speakers commonly leveraged in multimodal sentiment analysis. (3) Except for the modality gap that commonly presents in multimodal learning, there are additional semantic gaps for social media UGC, considering the fact that linguistic information in UGC focuses more on opinions reflecting sentiment polarities, while visuals imply more on the sentiment targets. These peculiarities limited the performance of methods developed for traditional opinion mining tasks [6], [7].

Please cite this article in press as: L. Zhang, L. Jin, G. Xu, X. Li, X. Sun, Z. Zhang, Y. Zhang, Q. Li. Optimal Target-Oriented Knowledge Transportation for Aspect-Based Multimodal Sentiment Analysis, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 4, pp. 59-69, 2025, <http://dx.doi.org/10.9781/ijimai.2024.02.005>

However, most existing research makes little targeted effort to social media aspect-based multimodal sentiment analysis. Traditional multimodal methods with early fusion in data-level [8], [9], or late fusion in decision-level [10], [11], suffer the problem of extra input redundancy and distributional modality gap, which result in the suboptimal performance for multimodal interaction. Although [1], [12], [13] have made attempts to eliminate the modality gap through modified cross-modal attention mechanisms, they neglect the semantic gap in social media user-generated content described above, especially the underlying target category knowledge in visual components. These semantic gaps may finally result in the increasing risk of misalignment in inter-modal interactions. Besides, previous works also neglect the importance of maintaining target sensitivity, which is particularly essential in acquiring discriminative multimodal representations to perform fine-grained analysis considering various aspect terms.

In this paper, we propose OtarNet, a multi-stage knowledge transportation framework based on optimal transport (OT) for ABMSA, which is effective in maintaining target sensitivity to avoid triviality and misalignment, caused by insufficient aspect interaction and semantic gaps respectively. Firstly, we explicitly transport the visual category knowledge through input space translation and reformulation, through which we acquire a synthetic sequence to supply context information. Secondly, the synthetic context sequence is incorporated into the unimodal encoding process, and ensures the good maintenance of target sensitivity through the proposed intra-modality target interaction mechanism, which outputs target-sensitive unimodal representations rich in semantic knowledge. Thirdly, the multiple unimodal representations are fed into the optimal transport interaction module, in which the inter-modality complementary knowledge (i.e. opinion knowledge in text and target knowledge in image) is implicitly transported to the other modality. Our contributions are summarized as follows:

- We propose OtarNet, a multi-stage knowledge transportation framework for aspect-based multimodal sentiment analysis. OtarNet explicitly transports the visual context knowledge before feature fusion to maintain target sensitivity, which is neglected by most multimodal approaches developed for traditional sentiment analysis. The proposed intra-modality target interaction mechanism is effective in avoiding triviality and misalignment.
- We leverage the optimal transport interaction to implicitly transport inter-modality complementary knowledge. OT interaction is effective in eliminating the distributional modality and semantic gap, which puts an extra burden on previous data-level or decision-level fusion techniques.
- We conduct extensive quantitative and qualitative experiments on three benchmark datasets: Twitter-15, Twitter-17 and Yelp. The newly-achieved state-of-the-art performance, together with the extensive ablation studies and visualizations demonstrate the superiority and effectiveness of OtarNet.

II. RELATED WORK

Despite the well-established field of sentiment analysis, our OtarNet focuses on aspect-based (aspect term) multimodal sentiment analysis, which is a novel challenge proposed firstly in 2019 by [3] and drawing increasing attention. This relatively new task stemmed from two lines of research, namely fine-grained sentiment analysis and multimodal sentiment analysis.

A. Fine-Grained Sentiment Analysis

Fine-grained sentiment analysis aims to identify the sentiment polarity of a textual sentence on a given aspect or target [14]. Its

research methods can be divided into three main groups: traditional feature selection based methods, neural network based methods and adaption of transformer-style models.

Early lexicon-based methods [15], [16] were established on handcrafted features such as lexical, syntactic and semantic features. These studies always demanded for a professional prior knowledge in linguistics [17], [18] and sometimes failed to capture the dependency between the given target and associated context. Later, neural networks with higher capability of encoding original features as continuous vectors were applied. [19]–[21] modified Long Short-Term Memory (LSTM) recurrent networks with stronger expressive power by attention mechanism to incorporate key information in sentence to a target aspect. [22] chose to use Gated Recurrent Unit (GRU) modules to utilize content information, which was able to deal with the syntactically structures of complex sentence. Moreover, sophisticated neural models with subtle intermediate attention were developed. [23] designed a Memory Network with multi-hop attention and external memory, which can explicitly capture the importance of each context word. [20], [24] leveraged multi-layer and multi-grained attention correspondingly to exploit semantic dependencies between opinion words in multi-level modeling for aspects. Recently, since the pre-trained language model [25] has made success in many tasks, [6] utilized BERT with an additional corpus and realized performance improvement in both aspect extraction and sentiment label classification. [26] achieved accurate prediction for this task which has been translated to a sentence-pair classification task by constructing auxiliary sentences. However, these studies fail to consider visual features that may boost these text-based approaches, which are one key factor of this paper.

B. Multimodal Sentiment Analysis

Multimodal sentiment analysis is an emerging research, the goal of which is to regress or classify the overall sentiment of an utterance integrating textual and non-textual information. Relative methods can also be divided into three groups: feature engineering methods, neural network based methods and modification of large pretrained models.

Early work mainly focused on feature engineering, which [27] combined adjective-noun pairs with linguistic features to calculate sentiment scores, and [28] proposed to fuse text and image features to obtain similarity of two instances for a new neighborhood classifier. Then motivated by the fusion approaches in feature and score-level [29], [30], pre-trained text and image CNNs [31] were conducted to extract feature and combine these multimodal features to train a logistics regression model. [32] modified LSTM to capture interactions between modalities through time. After that, models modified with attention mechanism [1], [12], [13] were proposed, [33] introduced a novel Attention-Based Modality-Gated Networks (AMGN) to learn the fine-grained correlation and the discriminative features between different modalities. In 2019, [3] introduced a Multi-Interactive Memory Network (MIMN) to supervise the textual and visual information under the given aspect. MIMN learned not only the interactive influences between cross-modality data but also the self influences in single-modality data. And more recently, [4] modified BERT architecture based on target-sensitive cross attention to capture the interaction between modalities. [34] proposed EF-CapTrBERT model to solve this task through input space translation, exploiting generated caption to substitute original images.

However, the existing approaches mainly focus on eliminating the modality gap generated by unimodal encoding procedures. Our OtarNet focuses more on target sensitivity and semantic gaps, which are solved based on multi-stage knowledge transportation and Optimal Transport Interaction.

C. Optimal Transport

Recently, Optimal Transport (OT) has attracted increasing attention in multiple fields [35]. As one of the research hotspots from optimization theory, OT has excellent performance on sequence alignment and domain adaption problems. By finding the best transportation plan between two data distributions with minimum cost, OT explicitly formulates signals to provide additional guidance [36]. Thus OT has achieved promising results compared to attention-based approaches guided by task-specific loss only [37]. For OT applications related to knowledge transportation, [38] explicitly distilled the knowledge of the monolingual summarization teacher into the student through an OT-based distance, which is effective in estimating the discrepancy and constructing cross-lingual correlation. And VOLT [39] formulated the quest of vocabularization as an optimal transport (OT) problem by finding the optimal transport matrix from the character distribution to the vocabulary token distribution. [40] used the transport plan as an ad-hoc attention score in the context of network embedding to align data modalities. MuLOT [41] utilized OT-based domain adaptation to learn strong cross-modal dependencies for sarcasm and humor detection. [42] innovatively revisited the label assignment from a global perspective and proposed to formulate the assigning procedure as an optimal transport (OT) problem. However, none of these studies have exploited optimal transport to implicitly incorporate complementary knowledge in ABMSA.

III. PROBLEM DEFINITION

Given a set of multimodal samples (e.g., tweets from Twitter) \mathcal{D} . Each piece of user-generated content $C \in \mathcal{D}$ consists of text information T with n words $[w_1, \dots, w_n]$ (e.g., [Taylor Swift drawn with colored pencils! emoji]) and an associated image I (e.g., first picture in Fig. 1). The sentiment target T_{tar} , as a sub-sequence of words in T is also given (e.g., [Taylor Swift]), which is assigned a sentiment label y_{tar} belonging to a given label set, such as {positive, negative, neutral} for Twitter and rating scores {1, 2, 3, 4, 5} for YELP. Our problem definition can be stated as follows: given \mathcal{D} as training corpus, the task goal is to learn a target-oriented sentiment classifier, so that it can

correctly predict sentiment labels y_{tar} for sentiment targets T_{tar} when encountering unseen samples. Note that there may be one or more targets mentioned in one sentence T , and the model needs to predict a single sentiment label for each associated target.






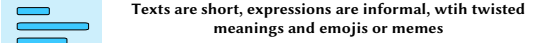
Aspect-Based Multimodal Sentiment Analysis				
Social Media Settings				
	Taylor Swift drawn with colored pencils!	British Library London	A lot of this going on in Chapel Hill, North Carolina today	Relaxing in a park of Alabama
Targets	Taylor Swift	British Library	Chapel Hill	Alabama
Visual Peculiarity				
Linguistic Peculiarity				

Fig. 1. An example of aspect-based multimodal sentiment analysis (ABMSA) from Twitter. One or more target aspect terms will be mentioned in the text for one piece of user-generated content (UGC). The difficulties of analyzing social media UGC lie in their convenient non-standard writing and network vocabulary.

IV. PROPOSED METHODOLOGY

In this section, we formulate our task firstly and then decompose OtarNet, as Fig. 2 shows, into three main components: (1) Intra-modality target knowledge transportation, which combines the semantic fusion and implicit fusion through incorporating constructed bridge sentence and bridge feature to calculate the target-sensitive feature. (2) Inter-modality complementary knowledge translation, which enhances the interactivity across image and text modalities. (3) Multimodal feature fusion, which conducts the final stage of fusion to capture the interior relationship between modalities. (4) Optimization process with a classifier is finally leveraged to minimize the standard

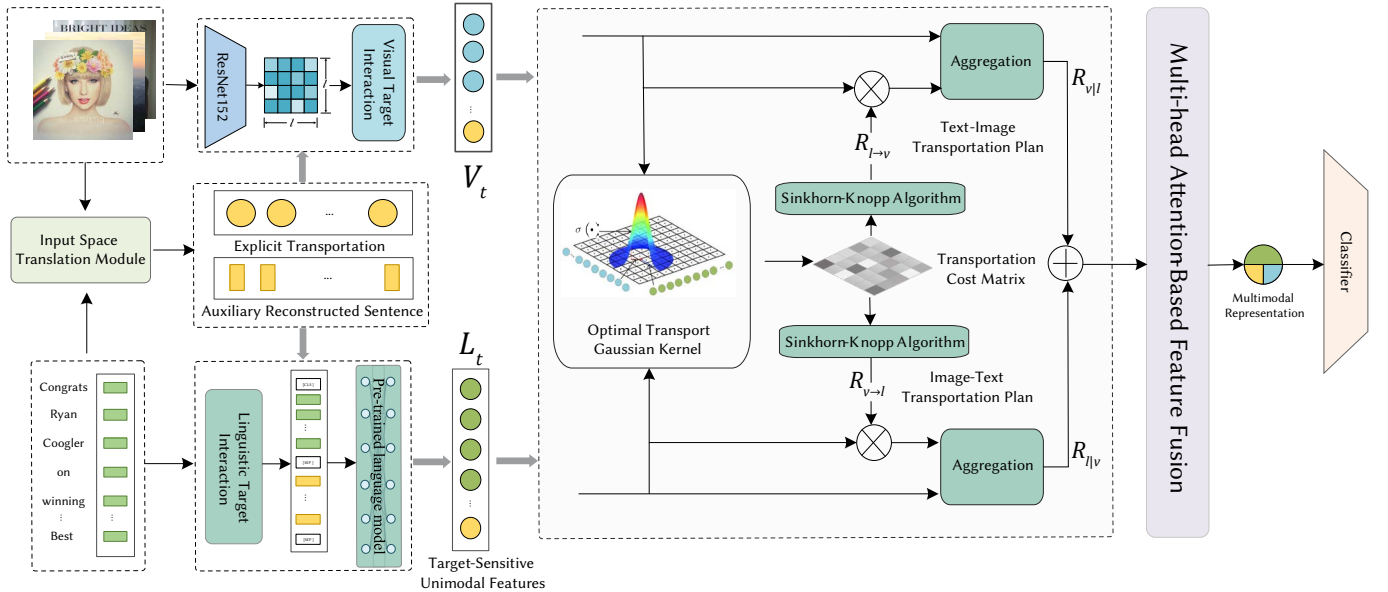


Fig. 2. The workflow of the proposed model OtarNet. Firstly, the auxiliary sentence with its representations are reconstructed through the Input Space Translation Module. Secondly, the auxiliary sequences containing target information are explicitly transported into the two unimodal encoding processes, which generate two target-sensitive features. Thirdly, the target-sensitive features are intergrated through the Optimal Transport-Based Interaction Module, which are designed to capture the complementary knowledge in multiple modalities. Finally, the multi-head attention-based fusion layer is leveraged to integrate the multimodal features, which are then used for classification.

cross-entropy loss function as the objective. The aforementioned modules are introduced in the following subsections, respectively.

A. Intra-Modality Target Knowledge Transportation

This section introduces the process of transporting target knowledge during the unimodal encoding procedure.

1. Input Space Translation & Reformulation

This module is leveraged to distill the object-level target information in complex visuals, and generate a synthetic context sequence with its features learned by a pretrained language model¹. The detailed information of this module is displayed in Fig. 3.

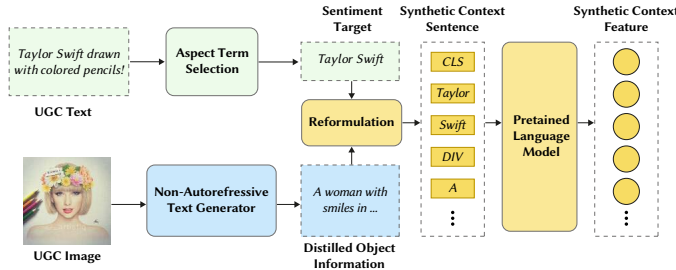


Fig. 3. Procedure of Input Space Translation & Reformulation.

Let C_u denote the content generated by user u , which contains a piece of text information (e.g. tweets, retweets or comments) T_u and an associated image I_u . To follow the work of [34], we exploit a transformed-style architecture to implement a non-autoregressive text generator $G_{Au}(I_u)$, through which the object-level target information is distilled into caption sentences. To leverage the aspect term information in the text modality, as well as the distilled visual information, we reconstruct the auxiliary sentence S_{Au} :

$$\begin{aligned} S_{Au} &= S_{trans}(I_u, T_u) \\ &= \{[CLS], [T_{as}], [DIV], [G_{Au}(I_u)]\} \\ &= \{[CLS], [t_1, \dots, t_k], [DIV], [c_1, \dots, c_m]\} \end{aligned} \quad (1)$$

where T_{as} denotes the aspect term containing a sub-sequence of k words $\{t_1, \dots, t_k\}$ from text information T_u , $G_{Au}(I_u)$ denotes the caption generator [34], implemented by utilizing a carefully designed variant model from DETR (DEtection TRansformer) [43] to get the m caption words $\{c_1, \dots, c_m\}$. Parameters of the caption generator G_{Au} in (1) are pretrained well and frozen during the whole experiment.

With the obtained auxiliary sentence, we can explicitly transport the context knowledge into the linguistic encoding process in the input space. And for the visual stream, the context knowledge can be transported in the feature space, thus we adopt a pretrained language model, which shares the same parameters with the linguistic encoder, to get the auxiliary context features F_{Au} .

2. Linguistic Knowledge Transportation & Encoding

This module is designed to incorporate target information while implementing the linguistic encoding process. As shown in Fig. 2, linguistic target interaction is leveraged before the encoding stage. As the pre-trained language models (like BERT, RoBERTa [6], [25]) can help acquire contextualized word representations with initialized parameters, which get well-trained over a large corpus. Thus the transformer-style encoders in sentence-pair classification mode are leveraged to integrate the input sentence and the reconstructed auxiliary sequence. The target-sensitive linguistic feature L_t can be achieved as:

$$L_t = LM([CLS], T_u, [SEP], S_{trans}(I_u, T_u)) \quad (2)$$

In (2), LM denotes the pretrained language models like BERT, RoBERTa, etc. '[CLS]' and '[SEP]' are the special tokens in the vocabulary used for classification and separation, and S_{trans} is the operation of input space translation introduced earlier in Fig. 2.

3. Visual Knowledge Transportation & Encoding

Dually, this module is designed to incorporate target information into visual feature space during the encoding process. Encoded visual features are firstly extracted from an input image I_u by ResNet [44]. The output size of the last convolutional layer in ResNet is $l \times l \times d_v$, where $l \times l$ denotes the l^2 block regions of an input image, d_v denotes the depth of feature map. The extracted visual feature of block regions $\{f^i\}_{i=1}^{l \times l}$ is fed into a linear transformation with matrix $W_v \in \mathbb{R}^{d_v \times d_h}$, in which d_h denotes the dimension of hidden states from BERT encoder. Thus the visual feature V is projected into the same space as the linguistic feature to match the embedding size of BERT:

$$\begin{aligned} V &= W_v \cdot \{f^i\}_{i=1}^{l \times l} \\ &= W_v \cdot \text{ResNet}(I_u) \end{aligned} \quad (3)$$

With the obtained auxiliary context features $F_{Au} \in \mathbb{R}^{N_2 \times d_h}$ and visual features V in (3), the target-sensitive visual representations can be achieved through visual target interaction, which conducts attentive interaction. Specifically, for the output of i -th head O_{head}^i we acquire the necessary query, key-value vectors through linear feature projection: $Q_b^i = W_q^i O_b$, $K_b^i = W_k^i V$, $V_b^i = W_v^i V$. The vectors are then used for calculating the attention output of the i -th head:

$$\begin{aligned} O_{head}^i &= \text{softmax}\left(\frac{Q_b^{iT} \cdot K_b^i}{\sqrt{d_h}}\right) V_b^i \\ &= \text{softmax}\left(\frac{[W_q^i O_b]^T \cdot [W_k^i V]}{\sqrt{d_h}}\right) W_v^i V \end{aligned} \quad (4)$$

All the attention outputs of m such heads $O_{head}^1, O_{head}^2, \dots, O_{head}^m$ in (4) are concatenated together, followed by a projection matrix W_j to get aggregated representation of m heads with residual connection and layer normalization (denoted as LN):

$$M_t = \text{LN}(V + W_j[O_{head}^1, O_{head}^2, \dots, O_{head}^m + b_j]) \quad (5)$$

Moreover, a dense layer and another residual connection are utilized from input V to the non-linear activated output feature of M_t in (5), followed by layer normalization to acquire the final target-sensitive visual feature V_t :

$$\begin{aligned} V_t &= V + \sigma(M_t + b_t) \\ &= V + \sigma(W_m \cdot W_j[O_{head}^1, O_{head}^2, \dots, O_{head}^m] + b_t) \end{aligned} \quad (6)$$

where $[\cdot]$ denotes the concatenation operation in feature dimension, σ is the non-linear activation function GELU [45], $W_j \in \mathbb{R}^{(m \times d_h) \times d_h}$, $W_m \in \mathbb{R}^{d_h \times d_h}$ are trainable parameters, V_t in (6) denotes target-sensitive visual feature, the final output of visual knowledge transportation.

B. Inter-Modality Complementary Knowledge Transportation

To enhance the interactivity across image and text modality, we exploit a recently proposed technique viz. optimal transport kernel (OTK) to incorporate information between heterogeneous modalities. OTK incorporates the idea of the optimal transport plan and kernel methods to fuse the unimodal features with varying dimensions and dependencies.

Let $V_t = (v_1, v_2, \dots, v_n)$ be the target-sensitive visual feature, $L_t = (l_1, l_2, \dots, l_p)$ denotes the target-sensitive linguistic feature obtained in (4) ($n = p$ is not necessary). Let κ be the Gaussian kernel with reproducing kernel Hilbert space (RKHS) \mathcal{H} and its associated kernel embedding $\varphi: \mathbb{R}^d \rightarrow \mathcal{H}$. Then we can get the $n \times p$ cost matrix K which carries the comparisons $\kappa(v_p, l_j)$ before alignment.

¹ <https://github.com/saahiluppall/catr/>

Then the transport plan between \mathbf{V}_t and \mathbf{L}_v denoted by the $n \times p$ matrix $\mathbf{P}(\mathbf{V}_t, \mathbf{L}_v)$ is defined as the unique solution of:

$$\min_{\mathbf{P} \in \mathcal{U}} \sum_{ij} \mathbf{C}_{ij} \mathbf{P}_{ij} - \varepsilon H(\mathbf{P}) \quad (7)$$

$$H(\mathbf{P}) = - \sum_{ij} \mathbf{P}_{ij} (\log(\mathbf{P}_{ij}) - 1) \quad (8)$$

where \mathbf{C}_{ij} in (7) represents the pairwise costs for aligning the elements of \mathbf{V}_t and \mathbf{L}_v . Equation (8) is the optimizing objective in the space of admissible couplings. To follow the recent work of [46], we choose $\mathbf{C} = -\mathbf{K}$ in our implementation, then the interaction based on transport matrix $\mathbf{P}(\mathbf{V}_t, \mathbf{L}_v)$ is defined as:

$$\begin{aligned} \Phi_l(v) &= \sqrt{p} \times \left(\sum_{i=1}^n \mathbf{P}_{i1} \varphi(v_i), \dots, \mathbf{P}_{ip} \varphi(v_i) \right) \\ &= \sqrt{p} \times \mathbf{P}(\mathbf{V}_t, \mathbf{L}_v)^T \varphi(v) \end{aligned} \quad (9)$$

The hybrid linguistic feature \mathbf{L}_h is obtained by aggregating the original target-sensitive linguistic feature \mathbf{L}_t and the visual transported interaction feature $\mathbf{R}_{v \rightarrow l} = \Phi_l(v)$ from (9):

$$\mathbf{L}_h = \mathbf{L}_t \oplus \mathbf{R}_{v \rightarrow l} = \mathbf{L}_t \oplus \Phi_l(v) \quad (10)$$

Similar to (10), we obtain the final hybrid visual feature \mathbf{V}_h through (11), which is also a result of concatenation:

$$\mathbf{V}_h = \mathbf{V}_t \oplus \mathbf{R}_{l \rightarrow v} = \mathbf{V}_t \oplus \Phi_v(l) \quad (11)$$

C. Multimodal Feature Fusion

Multimodal feature transfusion is designed to conduct the final stage of fusion, which captures the correlation between elements from different modalities. Through the above phase of OTI, two hybrid features \mathbf{V}_h and \mathbf{L}_h are obtained. OTI module adopts different vectors as queries to produce weighted representations, which are sensitive to features from different information streams. However, the features across modalities are involved in interactions through transportation weights, while the direct fusion of element values is still missing. So we propose to use multimodal feature transfusion based mainly on multi-head self-attention to capture the missing correlation in the element level. The input of multimodal feature transfusion is organized based on the two obtained hybrid features $\mathbf{V}_h, \mathbf{L}_h$ by OTI, and the target-sensitive visual feature \mathbf{V}_t .

The pooling operation is leveraged on visual features by taking the transformation of the first token. Then the output products are concatenated with linguistic features \mathbf{I}_m :

$$\mathbf{v}_{pooling} = \tanh(\mathbf{W}_p \cdot \mathbf{V}_t[1] + b_p) \quad (12)$$

$$\mathbf{v}'_{pooling} = \tanh(\mathbf{W}'_p \cdot \mathbf{V}_h[1] + b'_p) \quad (13)$$

$$\mathbf{I}_m = \mathbf{v}_{pooling} \oplus \mathbf{v}'_{pooling} \oplus \mathbf{L}_h \quad (14)$$

where $\mathbf{X}[1]$ in (12) and (13) denotes the first element of \mathbf{X} . And $\mathbf{I}_m \in \mathbb{R}^{(N_1+2) \times d_h}$ in (14) are fed into the multimodal feature transfusion module, which outputs \mathbf{O}_m as logits fed into the classifier.

D. Optimization Process

After the forward process of multimodal feature transfusion, we get the final multimodal hidden states \mathbf{O}_m for sentiment classification. Following previous work [4], [47], the pooled output of the first token is adopted, which is denoted as $\mathbf{H}_p \in \mathbb{R}^{d_h}$ and fed into a linear function followed by a softmax function for classification:

$$p(y|\mathbf{H}_p) = \text{softmax}(\mathbf{W}_c^t \cdot \text{dropout}(\mathbf{H}_p)) \quad (15)$$

In (15) $\mathbf{W}_c \in \mathbb{R}^{c \times d_h}$, c is the category number of dataset. All the

parameters in OtarNet are optimized through back propagation while minimizing the standard cross-entropy loss function defined in (16):

$$\mathcal{L} = - \frac{1}{|\mathcal{D}|} \sum_{k=1}^{|\mathcal{D}|} \log p(y^k | \mathbf{H}_p^k) \quad (16)$$

The overall training process is displayed in the following algorithm 1, in which line 1-6 initializes model parameters and input data, line 7-10 includes the process of input space translation based on input data, line 11-12 represents for the forward process of OtarNet to acquire final representation \mathbf{H}_p and line 14-17 refers to the optimization method in details.

Algorithm 1. Training Process of OtarNet

Input: Training Set \mathcal{D} , max number of epochs N_{epoch} , batch size β , learning rate η , parameters of caption generator θ_{cg} ,
Output: $\theta_{OtarNet}$
1: Initialize caption generator $CG(\theta_{cg})$
2: repeat
3: for $i = 1 \rightarrow \lfloor \frac{|\mathcal{D}|}{\beta} \rfloor$ do
4: $mini_batch \leftarrow \text{sample}(T, \beta)$
5: $L \leftarrow 0$
6: for $S \in mini_batch$ do
7: Forward image through encoder:
7: $\mathbf{V} \leftarrow \text{ResNet}(I)$
8: Forward image through caption generator:
8: $caption \leftarrow CG(I)$
9: Tokenize caption and tweet sentence S_t
10: Obtain S_{au} and F_{au} via the input space translation module:
10: $\{S_{au}, F_{au}\} \leftarrow MB(caption, S_t)$
11: Forward $\{S_t, S_{au}, F_{au}, V\}$ to get features:
11: $\mathbf{L}_t \leftarrow \text{BERT}(S_t + S_{au})$
11: $\mathbf{V}_t \leftarrow \text{ImplicitFusion}(\mathbf{V}, \mathbf{F}_{au})$
12: Forward $\{\mathbf{L}_t, \mathbf{V}_t\}$ to get the final \mathbf{H}_p
13: $\mathcal{L}(\mathbf{H}_p) \leftarrow - \frac{1}{|\mathcal{D}|} \sum_{k=1}^{|\mathcal{D}|} \log p(y^k | \mathbf{H}_p^k)$
14: $L \leftarrow L + \mathcal{L}(\mathbf{H}_p)$
15: end for
16: Update $\theta_{OtarNet}$ using $\Delta \mathcal{L}$
17: end for
18: until the evaluation results on the validation set drop continuously or this process has been iterated for N_{epoch} times

V. EXPERIMENTS

A. Datasets

We evaluate the OtarNet on three widely used benchmarks, including Twitter-15 [48], Twitter-17 [49] and Yelp². Their introduction are shown in Table I, with details displayed as follows:

TABLE I. DIFFERENCES OF BENCHMARKS

Datasets	Twitter-15	Twitter-17	Yelp
Data Source	Tweets in 2014-2015	Tweets in 2016-2017	Yelp Reviews
ContextSource	NA Text Generation	NA Text Generation	CrowdSourcing
Annotation Method	Nichesourcing	Nichesourcing	CrowdSourcing
Aspect Categories	Open Domain	Open Domain	Services
Class Number	3	3	5

² <https://www.yelp.com/dataset>

1. Twitter-15 and Twitter-17

These two sets consist of tweets including text and images posted during 2014-2015 and 2016-2017 respectively, whose sentiment labels over targets (i.e., entities in text), assigned from set {negative, neutral, positive} were supplemented later by [4]. The context information is collected by [34], which leveraged an object-aware transformer followed by a single-pass non-autoregressive text generation approach. The sentiment polarities toward each target were labeled by taking the majority label among three domain experts (Nichesourcing). The aspect categories contain various internet figures or events. Their statistics are displayed in Table II.

TABLE II. STATISTICS OF TWITTER DATASET

DataSet	Split	Positive	Neutral	Negative	Total
Twitter15	Training	928	1883	368	3179
	Validation	303	670	149	1122
	Test	317	607	113	1037
Twitter17	Training	1508	1638	416	3562
	Validation	515	517	144	1176
	Test	493	573	168	1234

2. Yelp

The third dataset we use is Yelp corpora obtained from Yelp Dataset Challenge. This corpora contains elaborate information on businesses across 10 cities, where we leverage complement reviews, photos, and corresponding captions. The sentiment polarities are labeled by directly taking the user ratings (from 1-5), and the task is a standard five-class classification problem. The evaluated categories are restricted to the provided services, such as their food, drink, environment, etc. Compared to Twitter sets, Yelp performs more fine-grained classification (5-class) based on well-organized reviews for specific domains. The statistics of Yelp are displayed in Table III.

TABLE III. STATISTICS OF YELP DATASET

Ratings	Training Set	Testing Set	Validation Set	Total
1	1248	384	387	2019
2	774	291	256	1321
3	1926	699	628	3253
4	504	164	177	845
5	1737	531	597	2865
Total	6189	2069	2045	10303

B. Evaluation Metrics

Following the previous work [4], [34], [50], we adopt Accuracy and Macro-F1 score (M-F1 for short) as our evaluation metrics. Accuracy can be calculated as follows:

$$\text{Accuracy} = \sum_{i=1}^C (TP_i + TN_i) / |D_{test}|$$

where C is the category numbers, $|D_{test}|$ is the total sample numbers in the test set. TP_i , TN_i are the numbers of True Positive and True Negative samples for the i -th category. To calculate Macro-F1, we firstly need to calculate the F1 score for each category based on their scores of precision and recall. The calculation of F1 for the i -th category $F1_i$ is defined as:

$$\text{Precision}_i = TP_i / (TP_i + FP_i)$$

$$\text{Recall}_i = TP_i / (TP_i + FN_i)$$

$$F1_i = \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

where TP_i , FP_i , FN_i , TN_i are the numbers of True Positive, False Positive, False Negative and True Negative samples for the i -th category. Based on the obtained F1 scores of each class, the Macro-F1 is defined as an average based on the categories:

$$\text{Macro-F1} = \sum_{i=1}^C F1_i / C$$

C. Experimental Settings

In our implementation, all the experiments are conducted with Pytorch on one 32G Tesla V100 GPU. We initialized pretrained weights of language models from HuggingFace³. As displayed in Table IV, the batch size is set as 32, and the maximum number of training epochs is set to 9. We apply the early stop strategy to avoid over-fitting. We train the models with an Adam weight decay optimizer with an initial learning rate of 5e-5. The optimal hyper-parameters are obtained by grid search. To ensure further reliability of our results and facilitate later explorations, we make our codes publicly available at <https://github.com/TomatoNLP/OTarNet>

For the input images, we adopt a pre-trained ResNet-152⁴, which outputs a feature map of size $7 \times 7 \times 2048$, indicating 49 block regions with depth d_v as 2048. For the input text, we maintain the standard configuration of BERT/RoBERTa and stack 12 BERT layers. The feature dimension of hidden state output by one BERT layer dh is 768, which is calculated by inner multi-head attention with $m = 12$ heads. We then truncate the max input length N_1 and the max bridge sentence length N_2 to 125. Besides, the other settings of hyperparameters during the training process are displayed in Table IV.

TABLE IV. HYPERPARAMETER SETTINGS

Hyperparameter	Symbol	Value
Epochs	E	9
Batchsize	B	32
Dropout	d	0.15
Learning Rate	lr	5e-5
OTI layer	LOTI	1
ME Layer	LME	1
Weight Decay	Wd	0.01
Optimizer	-	AdamW
BERT Weights	-	bert-base-uncased
RoBERTa Weights	-	bertweet-base

D. Model Zoo

In this subsection, we will give comprehensive introductions to the leveraged baseline models, including:

- **EF-Net**: An attention capsule extraction and multi-head fusion network for MABSA, which is established based on multi-head attention (MHA) and the ResNet-152
- **Res-MGAN**: A combination of textual and visual contents from ResNet and MGAN. It is implemented by concatenating the pooling results of ResNet and MGAN, which is a multi-grained attention network proposed in [51] for fusing the target and the context.
- **Res-BERT + BL**: Similar to Res-MGAN, Res-BERT + BL is a combination of textual and visual content from ResNet and BERT. BL denotes another BERT layer on the top, which is leveraged for feature fusion.

³ <https://huggingface.co/models>

⁴ <https://download.pytorch.org/models/resnet152-b121ed2d.pth>

- **mPBERT**: A variant of mBERT, which uses the max pooling of visual features and first token pooling ([CLS]) to obtain the final output.
- **RelConsTransLG**: A constituent-based transformer, which applies meta auxiliary learning to generate labels on edges between tokens, and can induce constituents without constituent parsers for MABSA.
- **TomBERT**⁵: A multimodal backbone leveraging a target attention mechanism to perform target-image matching, which is helpful for deriving target-sensitive visual representations.
- **EF-CapTrBERT**⁶: A two-stream multimodal backbone, which leverages space translation to construct an auxiliary sentence for language models.

E. Overall Performance

We compare OtarNet against a collection of neural network based or modified transformer models designed for multimodal aspect-based sentiment analysis. The model performance on three benchmark datasets is displayed in Table V, in which the Twitter results of compared models are directly quoted from published articles, and the Yelp results are obtained through our reimplementation [52].

Based on the displayed experimental results we can make a couple of observations: (1) Our model OtarNet outperforms former multimodal models and achieve the new SOTA performance on three benchmark datasets, demonstrating the effectiveness of our work. (2) Compared to the second-best model, our OtarNet further enhance the performance by a margin and achieve an average of 2.91%, 5.56%, and 2.04% performance improvement respectively on the three datasets from the perspective of Macro-F1, which further indicates the effectiveness of our framework. (3) Since the corpora of Twitter datasets inevitably contains noisy UGC on Twitter websites, the model performance is relatively lower than those in well organized Yelp dataset.

F. Further Analysis

The previous SOTA model EF-CapTrBERT is sub-optimal, which leverages distilled visual knowledge to perform the early fusion. The degradation may come from the noise during distillation and the lack

of complementary details from the visual stream. Thus EF-CapTrBERT finally achieved competitive results to a variation of our model OtarNet + T-Trans, which adopt similar settings to merely transport target knowledge to the text modality. And another competitive baseline TomBERT leverage cross-model attention mechanisms to realize target-sensitive visual representations. Their methods neglect the process of maintaining target sensitivity in linguistic encoding, which is a main difference that leads to limited performance. Compared to previous SOTA TomBERT, EF-CapTrBERT, and our variations, the performance improvements of OtarNet + M-Trans confirm the achievements of our goals, including maintaining target sensitivity and complementary knowledge transportation, which get further verification and analysis in the following ablations and visualizations.

G. Effectiveness of Target Knowledge Transportation

To achieve the objective of transporting target category knowledge in visual components, we exploit a transformer-style architecture to distill the object-level target information in complex visuals. The overall procedure is displayed in Fig. 3. Contrastively, we conduct ablation experiments on three test sets by decomposing the Input Space Translation Module. Specifically, different transportation plans are closely attempted, including T-Trans, I-Trans and M-Trans, which means transporting the target information into different modalities (Text, Image, Multimodal) to explore the effectiveness of bidirectional target knowledge transportation.

As shown in Table V, all the settings of multimodal data with different transportation settings achieve better performance compared to those without target knowledge, which can well confirm the bridge effect in fusing information from multiple modalities. For the transportation plans of target knowledge, M-Trans achieves expected predominant results compared to models with other settings, which turns out that bidirectional integration can achieve better accomplishment for the goal of incorporating target knowledge, and function better bridge effect compared to unimodal transportation plans.

To provide a more intuitive comparison, we provide attention maps in Fig. 4 for the weights in the matrix of optimal transportation plans. With the transportation of distilled visual information (b), OtarNet pays more attention to the sentiment target in visuals compared to (a). The comparison results are explicitly met with our goal of maintaining target sensitivity.

TABLE V. OVERALL PERFORMANCE ON TWO TWITTER DATASETS AND YELP

Comparisons	Model	Twitter-15		Twitter-17		Yelp	
		Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
Baselines	Res-MGAN [4]	71.65	63.88	66.37	63.04	80.62	71.37
	EF-Net [50]	73.65	67.90	67.77	65.32	81.33	71.50
	Res-BERT+BL [4]	75.02	69.21	69.20	66.48	80.93	71.75
	mPBERT(CLS) [48]	75.79	71.07	68.80	67.06	79.81	68.52
	RelConsTransLG [53]	76.80	73.30	69.80	68.50	80.15	68.25
	TomBERT(FIRST) [4]	77.25	71.75	70.34	68.03	81.46	73.44
	EF-CapTrBERT [34]	78.35	73.61	69.93	68.90	82.14	74.15
Ours	OtarNet - TG	74.67	69.33	68.34	67.52	78.87	71.67
	OtarNet + T-Trans	78.23	72.94	72.54	70.68	81.67	74.68
	OtarNet + I-Trans	76.48	70.33	69.68	68.27	80.85	73.11
	OtarNet + M-Trans	80.63	76.32	74.57	72.73	84.83	76.66
Margin	$\delta_{\text{ours-second_best}}$	$\Delta 2.28$	$\Delta 2.71$	$\Delta 4.23$	$\Delta 3.83$	$\Delta 2.69$	$\Delta 1.51$

The numbers in **bold face** denotes the best results, The numbers in **bold face** denotes the best results in Baselines
TG denotes the Target Knowledge obtained through the input space translation module.

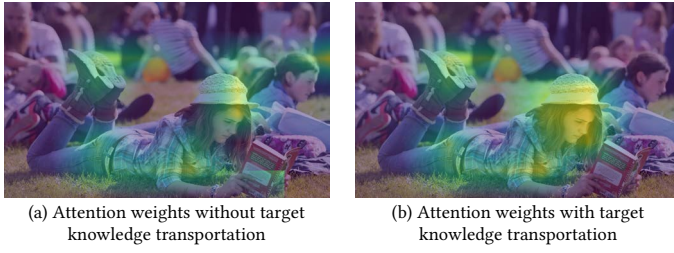


Fig. 4. In this sample, the distilled target knowledge through **S_trans** in (1) is: **A girl sitting on the ground with a baseball bat**. As the attention map in (a), the visual features without target knowledge are less discriminative. By integrating the distilled context information, Otanet captures better visual semantics as shown in (b). More visualizations are provided in the later section of **Qualitative Analysis**.

H. Effectiveness of Optimal Transport Interaction

We decompose the optimal transport interaction into two single parts, including Image_to_Text transportation (IOT) and Text_to_Image transportation (TOI), then add one or both of them to Otanet to demonstrate the effectiveness of transport modules. To make a further step, we explore two ways of input to produce different query vectors for transport plans, which further explains the combined effect of the target knowledge and optimal transport interaction.

Depicted by the results of accuracy (ACC.) and Macro-F1 (F1) in Table VI, outputs with bidirectional optimal transport interaction achieve the best performance improvement due to the bidirectional complementary knowledge transportation. The superiority of Otanet + Bi-CKT indicates that, the proposed Optimal Transport Interaction method succeeds in transporting the inter-modal complementary knowledge, and produces hybrid features with more comprehensive information for analyzing targets' sentiment. Besides, we also perform an ablation study and deepen the block architecture by removing or stacking the same transport layer, as Fig. 5 shows. The ablation of the OTI layer naturally brings performance degradation. Nevertheless, we observe little or tiny performance boost with a deeper interaction module, indicating that two layers of optimal transport interaction are sufficient for transporting complementary knowledge. This may be because the caption carrying part of the target information, as enhanced image attributes, has interacted with linguistic information through input space translation. [34].

TABLE VI. EFFECTIVENESS OF OPTIMAL TRANSPORT INTERACTION

Datasets & Methods	Twitter-15		Twitter-17		Yelp	
	ACC.	F1	ACC.	F1	ACC.	F1
Otanet - CKT	76.82	73.74	72.01	69.84	83.24	75.53
Otanet + TOI-CKT	78.25	74.32	72.37	71.25	83.75	76.14
Otanet + IOT-CKT	78.73	74.57	72.98	71.65	83.55	75.74
Otanet + Bi-CKT	80.63	76.32	74.57	72.73	84.83	76.66

CKT is the abbreviation for complementary knowledge transportation. TOI-CKT refers to Text_to_Image knowledge transportation, IOT-CKT refers to Image_to_Text transportation, and Bi-CKT refers to the bidirectional optimal transport interaction.

I. Effectiveness of Multimodal Feature Transfusion

In this section, we conduct two main experiments on the Yelp dataset for the multi-head attention-based fusion (MHAF) module to validate the effectiveness and find the best settings. We ablate the MHAF module and feed multiple types of features to the classifier. The results are shown in Fig. 6.

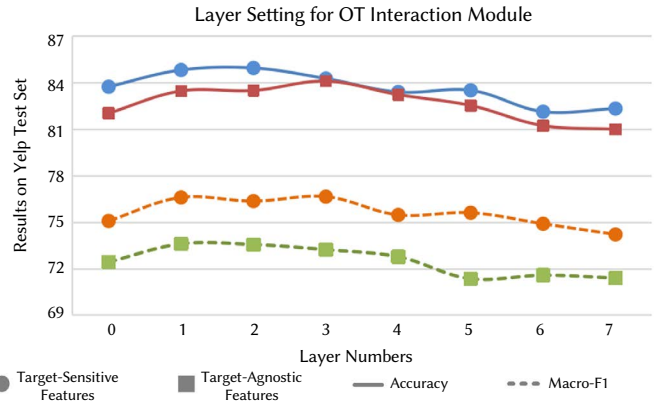


Fig. 5. Layer settings of optimal transport interaction module and results on Yelp test set.

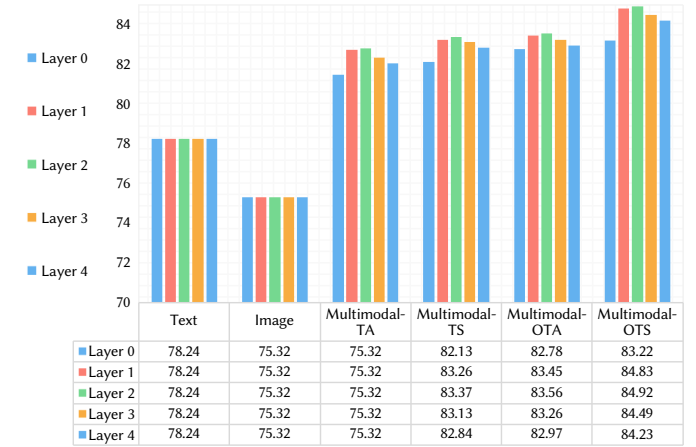


Fig. 6. Ablation studies for multimodal feature fusion layer. The accuracy results on the Yelp test set are displayed. TA denotes the Target-Agnostic feature, acquired by base encoders without target-oriented knowledge transportation. TS denotes the Target-Sensitive feature, acquired by the target interaction modules without complementary knowledge transportation. And OTA, OTS are the TA, TS features after OT interaction.

Conclusions can be made that the Multimodal-OTS feature is definitely more representative to provide discriminative details for classification. The multi-head attention-based feature fusion (MHAF) module effectively receives an average performance boost of 2.90% due to the inner interactive mechanism. To make a further exploration for the depth of the MHAF module, we add the inner attention layers to conduct another experiment, the results of which are shown in Fig. 6. However, as the same result in OTI, we find the model performance drops slightly or grows at a very slow pace.

J. Qualitative Analysis

In this section, we present some examples from trained models to provide several qualitative analyses, including a case study and attention maps to better understand what Otanet has learned.

1. Case Study

Fig. 7 shows two predictions with the text attention visualizations of Otanet. The displayed representative samples confirm the peculiarity that images focus more on targets while sentences express opinions. This peculiarity emphasizes the necessity of transporting complementary knowledge in different modalities of information. This objective is achieved by Otanet through the Optimal Transport Kernel method, and can be reflected by text attention visualizations, in which key opinion words in sentences are successfully captured. Results are obtained through Text-BERT, Multimodal-BERT, and

our OtarNet, which incorporate different input information. For the confusing information (e.g., 'Best Independent Film for Fruitvale') in unimodal, Text-BERT fails to leverage the combined effect from multiple sources and Multimodal-BERT may misunderstand the key target in visuals. However, our OtarNet with multiple knowledge transportation leveraging context information succeeds to catch the key differences and make accurate predictions.



Inputs		
		
Congrats Ryan Coogler (1) on winning Best Independent Film for Fruitvale at @theaafca !		Robert Downey Jr (1) . paid a visit to the Great Ormond Street Hospital (2) in London and met young super fans
Text Attention	[CLS] Congrats Ryan Coogler on winning Best Independent Film for Fruitvale at @theaafca [SEP] A man in a suit and tie standing in front of a wall	[CLS] Robert Downey Jr paid a visit to the Great Ormond Street Hospital in London and met young super fans [SEP] A man and a little girl standing next to each other
Ground Truth	(1)-POS, (2)-NEU	(1)-POS, (2)-NEU
Text-Bert	(1)-POS ✓, (2)-NEU ✗	(1)-NEU ✗, (2)-NEU ✓
Multimodal-Bert	(1)-POS ✓, (2)-POS ✗	(1)-NEU ✗, (2)-NEU ✓
OtarNet	(1)-POS ✓, (2)-NEU ✓	(1)-POS ✓, (2)-NEU ✓

Fig. 7. Examples that text-only or classic multimodal BERT with text and image make the wrong predictions, but our proposed approach with context information insertion gets correct. The text attention indicates the importance of different words computed by the OtarNet.

2. Visualizations

In order to further validate the combined effect between images and text, we visualize the obtained attention weights of block regions in the visual feature map as shown in Fig. 8. And it is obvious that some key factors in pictures like emotions or actions are necessary to infer the implicit sentiment in text. Specifically, for the target 'NFL' in the first example, it's nearly impossible to infer the implied negative orientation based only on text modality. Nevertheless, with the action of clutching her chest and the pained expression on her face in the picture, our OtarNet can make the correct prediction. Similar examples are displayed in Fig. 8, which demonstrates that our OtarNet has the ability to capture key details in visual features helpful for judgment.

VI. CONCLUSION

This paper proposes a novel OtarNet for multimodal aspect-based sentiment analysis. Different from previous works, which are suffering from the problems of lacking target interaction and distributional modality gap, our OtarNet leverage multi-stage interaction mechanisms to transport knowledge from multiple perspectives for solving the issues above. To maintain interactions with aspect terms for target sensitivity, we leverage an input space translation and multistage interaction method to capture the intra-modality target knowledge of social media content. To capture the inter-modality complementary knowledge, OtarNet exploits a novel approach of the Optimal Transport Kernel method. Compared to attention mechanisms guided by task-specific loss only, OtarNet based on Optimal Transport offers additional signals by reformulating the multimodal fusion as a transportation problem. Experiments on three real-world datasets demonstrate the effectiveness and superiority of our model, which gets further indicated in the ablation study and visualizations. In general, OtarNet exhibits excellent effectiveness in the fine-grained sentiment analysis of open-domain social media content and cross-modal complementation. In the future, we will consider designing models with other ways of interaction, including graph aggregation or loss regulations. We also plan to apply our model to solve related problems such as fine-grained multimodal aligning in hate detection,















Sentiment Target Positive Negative	Input Text	Generated Caption	Input Image	Attention visualization
NFL	When a guy you used to talk to got drafted to the NFL	A woman sitting on a bench clutching her chest		
Cannes	Might take a flight out for the weekend, and go Cannes - cause I can	A woman is jumping high on the beach		
Frankie Dettori	Frank Dettori wins the Dante Stakes on Wings of Desire! What a week this man is having!	A man wearing a helmet and sunglasses and a baseball cap		
Vince Gilligan	Happy birthday, Vince Gilligan! Happy birthday, master!	A man smiling in a blue jacket		
Confederation of Progressive Trade Unions	Uproar in front of Confederation of Progressive Trade Unions	A woman crying on the ground with a man and a woman by her side		
Eric Church	Eric Church Will Rock the Taste of Country Music Festival as 2018 Headliner!	A man in a black shirt and a guitar and microphones		
BellaRusso 14	♥️ can't wait for another summer of concerts and fun w beano bag @ BellaRusso14	A woman carrying a little girl sitting on the grass		

Fig. 8. Attention map of several examples in the Twitter dataset. The auxiliary sentences provided by the non-autoregressive text generator are also provided. We select the top-K values in the optimal transportation plans to make visualizations, which can partly reflect what the OtarNet has learned to integrate. The displayed results confirm the model's capacity of capturing key details in visual features.

as well as multi-lingual applications under low-resource language scenarios. We are also interested in incorporating data from other sources such as speeches or videos.

ACKNOWLEDGEMENT

The work is supported by the National Natural Science Foundation of China (62206267).

REFERENCES

- [1] F. Huang, X. Zhang, Z. Zhao, J. Xu, Z. Li, "Image- text sentiment analysis via deep multimodal attentive fusion," *Knowledge-Based Systems*, vol. 167, pp. 26–37, 2019.
- [2] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowledge-based systems*, vol. 161, pp. 124–133, 2018.
- [3] N. Xu, W. Mao, G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 371–378, AAAI Press.
- [4] J. YU, J. JIANG, "Adapting bert for target-oriented multimodal sentiment classification," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019, pp. 5408–5414.
- [5] M. E. Basiri, M. Abdar, M. A. Cifci, S. Nemati, U. R. Acharya, "A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques," *Knowledge-Based Systems*, vol. 198, p. 105949, 2020.
- [6] H. Xu, B. Liu, L. Shu, S. Y. Philip, "Bert post-training for review reading comprehension and aspect-based sentiment analysis," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 2324–2335.

- [7] M. Li, L. Chen, J. Zhao, Q. Li, "Sentiment analysis of chinese stock reviews based on bert model," *Applied Intelligence*, vol. 51, no. 7, pp. 5016–5024, 2021.
- [8] V. Pérez-Rosas, R. Mihalcea, L.-P. Morency, "Utterance- level multimodal sentiment analysis," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013, pp. 973–982.
- [9] V. P. Rosas, R. Mihalcea, L.-P. Morency, "Multimodal sentiment analysis of spanish online videos," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 38–45, 2013.
- [10] F. Celli, B. Lepri, J.-I. Biel, D. Gatica-Perez, G. Riccardi, F. Pianesi, "The workshop on computational personality recognition 2014," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1245–1246.
- [11] J. G. Ellis, B. Jou, S.-F. Chang, "Why we watch the news: a dataset for exploring sentiment in broadcast video news," in *Proceedings of the 16th international conference on multimodal interaction*, 2014, pp. 104–111.
- [12] X. Yang, S. Feng, D. Wang, Y. Zhang, "Image- text multimodal emotion classification via multi-view attentional network," *IEEE Transactions on Multimedia*, vol. 23, pp. 4014–4026, 2021.
- [13] F. Chen, Z. Yuan, Y. Huang, "Multi-source data fusion for aspect-level sentiment classification," *Knowledge- Based Systems*, vol. 187, p. 104831, 2020.
- [14] W. An, F. Tian, P. Chen, Q. Zheng, "Aspect-based sentiment analysis with heterogeneous graph neural network," *IEEE Transactions on Computational Social Systems*, 2022.
- [15] J. Wagner, P. Arora, S. Cortes, U. Barman, D. Bogdanova, J. Foster, L. Tounsi, "Dcu: Aspect- based polarity classification for semeval task 4," *SemEval 2014*, p. 223, 2014.
- [16] B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 2002, pp. 79–86.
- [17] M. S. Akhtar, D. Gupta, A. Ekbal, P. Bhattacharyya, "Feature selection and ensemble construction: A two- step method for aspect based sentiment analysis," *Knowledge-Based Systems*, vol. 125, pp. 116–135, 2017.
- [18] L. Jiang, M. Yu, M. Zhou, X. Liu, T. Zhao, "Target- dependent twitter sentiment classification," in *Proceedings of the 49th annual meeting of the association for computational linguistics*, 2011, pp. 151–160.
- [19] N. Liu, B. Shen, "Aspect-based sentiment analysis with gated alternate neural network," *Knowledge-Based Systems*, vol. 188, p. 105010, 2020.
- [20] P. Chen, Z. Sun, L. Bing, W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 452–461.
- [21] Y. Wang, M. Huang, X. Zhu, L. Zhao, "Attention- based lstm for aspect-level sentiment classification," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 606–615.
- [22] Q. Liu, H. Zhang, Y. Zeng, Z. Huang, Z. Wu, "Content attention model for aspect based sentiment analysis," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 1023–1032.
- [23] D. Tang, B. Qin, T. Liu, "Aspect level sentiment classification with deep memory network," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 214–224.
- [24] F. Fan, Y. Feng, D. Zhao, "Multi-grained attention network for aspect-level sentiment classification," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 3433–3442.
- [25] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 4171–4186.
- [26] C. Sun, L. Huang, X. Qiu, "Utilizing bert for aspect- based sentiment analysis via constructing auxiliary sentence," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 380–385.
- [27] D. Borth, R. Ji, T. Chen, T. Breuel, S.-F. Chang, "Large- scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 223–232.
- [28] Y. Zhang, L. Shang, X. Jia, "Sentiment analysis on microblogging by integrating text and image features," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2015, pp. 52–63, Springer.
- [29] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 163–171.
- [30] H. Wang, A. Meghawat, L.-P. Morency, E. P. Xing, "Select-additive learning: Improving generalization in multimodal sentiment analysis," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 2017, pp. 949–954, IEEE.
- [31] Y. Yu, H. Lin, J. Meng, Z. Zhao, "Visual and textual sentiment analysis of a microblog using deep convolutional neural networks," *Algorithms*, vol. 9, no. 2, p. 41, 2016.
- [32] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, L. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 5642–5649, AAAI Press.
- [33] F. Huang, K. Wei, J. Weng, Z. Li, "Attention-based modality-gated networks for image-text sentiment analysis," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 3, pp. 1–19, 2020.
- [34] Z. Khan, Y. Fu, "Exploiting bert for multimodal target sentiment classification through input space translation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3034–3042.
- [35] H.-Y. Lin, H.-H. Tseng, X. Lu, Y. Tsao, "Unsupervised noise adaptive speech enhancement by discriminator- constrained optimal transport," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19935–19946, 2021.
- [36] L. Chen, Z. Gan, Y. Cheng, L. Li, L. Carin, J. Liu, "Graph optimal transport for cross-domain alignment," in *International Conference on Machine Learning*, 2020, pp. 1542–1553.
- [37] E. Grave, A. Joulin, Q. Berthet, "Unsupervised alignment of embeddings with wasserstein procrustes," in *The 22nd International Conference on Artificial Intelligence and Statistics*, vol. 89, 2019, pp. 1880–1890.
- [38] T. T. Nguyen, A. T. Luu, "Improving neural cross- lingual abstractive summarization via employing optimal transport distance for knowledge distillation," in *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, 2022, pp. 11103–11111, AAAI Press.
- [39] J. Xu, H. Zhou, C. Gan, Z. Zheng, L. Li, "Vocabulary learning via optimal transport for neural machine translation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021, pp. 7361–7373.
- [40] L. Chen, G. Wang, C. Tao, D. Shen, P. Cheng, X. Zhang, W. Wang, Y. Zhang, L. Carin, "Improving textual network embedding with global attention via optimal transport," in *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 2019, pp. 5193–5202.
- [41] S. Pramanick, A. Roy, V. M. Patel, "Multimodal learning using optimal transport for sarcasm and humor detection," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 546–556.
- [42] Z. Ge, S. Liu, Z. Li, O. Yoshie, J. Sun, "OTA: optimal transport assignment for object detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 303–312.
- [43] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, "End-to-end object detection with transformers," in *16th European Conference on Computer Vision*, vol. 12346, 2020, pp. 213–229, Springer.
- [44] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [45] D. Hendrycks, K. Gimpel, "Bridging nonlinearities and stochastic regularizers with gaussian error linear units," *CoRR*, vol. abs/1606.08415, 2016.
- [46] G. Mialon, D. Chen, A. d'Aspremont, J. Mairal, "A trainable optimal transport embedding for feature aggregation and its relationship to attention," in *ICLR 2021-The Ninth International Conference on Learning Representations*, 2021.
- [47] X. Li, L. Bing, W. Zhang, W. Lam, "Exploiting BERT for end-to-end aspect-based sentiment analysis," in *Proceedings of the 5th Workshop on Noisy User-generated Text*, 2019, pp. 34–41.
- [48] D. Lu, L. Neves, V. Carvalho, N. Zhang, H. Ji, "Visual attention model for name tagging in multimodal social media," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 1990–1999.

- [49] Q. Zhang, J. Fu, X. Liu, X. Huang, "Adaptive co-attention network for named entity recognition in tweets," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, vol. 32, 2018, AAAI Press.
- [50] D. Gu, J. Wang, S. Cai, C. Yang, Z. Song, H. Zhao, L. Xiao, H. Wang, "Targeted aspect-based multimodal sentiment analysis: An attention capsule extraction and multi-head fusion network," *IEEE Access*, vol. 9, pp. 157329–157336, 2021.
- [51] F. Fan, Y. Feng, D. Zhao, "Multi-grained attention network for aspect-level sentiment classification," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3433–3442.
- [52] D. Q. Nguyen, T. Vu, A. T. Nguyen, "Bertweet: A pre-trained language model for english tweets," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 9–14.
- [53] L. M. S. Khoo, H. L. Chieu, "Meta auxiliary labels with constituent-based transformer for aspect-based sentiment analysis," 2020.



Linhao Zhang

Linhao Zhang received a B.S. degree from Xidian University, Xi'an, China, in 2020. He is working towards the Ph.D. degree in Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include affective computing and multimodal learning.



Li Jin

Li Jin received the B.S. degree from Xidian University, Xi'an, China, in 2012 and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2017. He is currently an Associate Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include machine learning,

knowledge graph and geographic information processing.



Guangluan Xu

Guangluan Xu received the B.Sc. degree from Beijing Information Science and Technology University, Beijing, China, in 2000, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2005. He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research

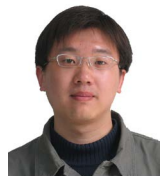
interests include data mining and machine learning.



Xiaoyu Li

Xiaoyu Li received the B.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2016 and M.E. degree from Beijing University of Posts and Telecommunications in 2019. He is currently a Research Assistant Fellow at the Aerospace Information Innovation Institute, Chinese Academy of Science, Beijing, China.

His research interests include data mining, information extraction, event logic graph and natural language processing.



Xian Sun

Xian Sun received the B.Sc. degree from Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004. He received the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2009. He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research

interests include computer vision, geospatial data mining, and remote sensing image understanding.



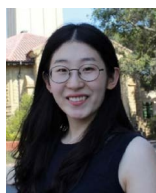
Zequn Zhang

Zequn Zhang received a B.Sc. degree from Peking University, Beijing, China, in 2012, and a Ph.D. degree from Peking University in 2017. He is currently a Research Assistant at the Aerospace Information Innovation Institute, Chinese Academy of Science, Beijing, China. His research interests include information fusion knowledge graphs and natural language processing.



Yanan Zhang

Yanan Zhang received the B.S. degree in communication engineering from Sichuan University, Chengdu, China, in 2016, and the Ph.D. degree from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2022. She is currently an Assistant Professor with Sichuan University. Her research interests include multimodal sentiment analysis and question answering.



Qi Li

Qi Li received the B.S. degree from Dalian University of Technology, Dalian, China, in 2013 and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2018. She is currently a Senior Engineer in the Faculty of Psychology, Beijing Normal University. Her research interests include computational psychology and psychoanalysis in social networks.

TKU-PSO: An Efficient Particle Swarm Optimization Model for Top-K High-Utility Itemset Mining

Simen Carstensen¹, Jerry Chun-Wei Lin² *

¹ University of Bergen, Bergen, (Norway)

² Western Norway University of Applied Sciences, Bergen, (Norway)

* Corresponding author: simencarstensen@gmail.com (S. Carstensen), jerrylin@ieee.org (J. C.-W. Lin).

Received 16 September 2022 | Accepted 11 November 2022 | Early Access 15 January 2024



ABSTRACT

Top-k high-utility itemset mining (top-k HUIM) is a data mining procedure used to identify the most valuable patterns within transactional data. Although many algorithms are proposed for this purpose, they require substantial execution times when the search space is vast. For this reason, several meta-heuristic models have been applied in similar utility mining problems, particularly evolutionary computation (EC). These algorithms are beneficial as they can find optimal solutions without exploring the search space exhaustively. However, there are currently no evolutionary heuristics available for top-k HUIM. This paper addresses this issue by proposing an EC-based particle swarm optimization model for top-k HUIM, which we call TKU-PSO. In addition, we have developed several strategies to relieve the computational complexity throughout the algorithm. First, redundant and unnecessary candidate evaluations are avoided by utilizing explored solutions and estimating itemset utilities. Second, unpromising items are pruned during execution based on a threshold-raising concept we call minimum solution fitness. Finally, the traditional population initialization approach is revised to improve the model's ability to find optimal solutions in huge search spaces. Our results show that TKU-PSO is faster than state-of-the-art competitors in all datasets tested. Most notably, existing algorithms could not complete certain experiments due to excessive runtimes, whereas our model discovered the correct solutions within seconds. Moreover, TKU-PSO achieved an overall accuracy of 99.8% compared to 16.5% with the current heuristic approach, while memory usage was the smallest in % of all tests.

KEYWORDS

Data Mining, Evolutionary Computation, Fitness Estimation, Particle Swarm Optimization, Threshold-Raising Strategy, Top-k High-Utility Itemset.

DOI: 10.9781/ijimai.2024.01.002

I. INTRODUCTION

DATA mining is a popular field of research focused on extracting interesting patterns from massive datasets. These patterns are highly beneficial as they can help reveal and comprehend hidden relationships within data. Several distinctive data mining approaches exist, each specialized in locating a specific type of pattern.

Frequent itemset mining (FIM) [1] is a subfield within data mining for finding item combinations (itemsets) that occur no less than a minimum support count, where the support describes the number of transactions that contain the itemset. In other words, FIM returns the most prevalent patterns in the data. There is a wide variety of applications for FIM, such as finding co-occurring words in a text or products often bought together in a store. However, the usefulness of FIM is limited as it assumes frequency always defines itemset importance. Concerning customer purchases, businesses are typically interested in the patterns that contribute the most profit, and these itemsets are not necessarily among the common purchases. For this reason, data mining based on utilities has been proposed.

High-utility itemset mining (HUIM) [2] is an extension of FIM for discovering valuable patterns within data. The value of an itemset is quantified by a utility, and HUIM algorithms aim to reveal all itemsets with utility over a user-specified minimum utility threshold (HUIs). A key property of this strategy is that the utility can characterize different quality measures of itemsets, e.g., profit, cost, time, or even frequency. This way, HUIs can fit a wider variety of analytical problems than the frequent patterns produced by FIM. The most common application of HUIM is to identify consumer behaviors through market basket analysis [3]. However, recent studies have also shown its usefulness in problems such as emerging topic detection [4], travel pattern analysis [5], and cardiovascular disease detection [6].

Although there has been extensive research on HUIM, the algorithms tend to be unintuitive in practice. The required minimum utility threshold is challenging to set properly without knowing specific data characteristics. Typically, the user has to test multiple threshold values to find a reasonable number of patterns, which may not be feasible depending on the model's runtime. Top-k HUIM

Please cite this article in press as:

S. Carstensen, J. Chun-Wei Lin. TKU-PSO: An Efficient Particle Swarm Optimization Model for Top-k High-Utility Itemset Mining, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 4, pp. 70-81, 2025, <http://dx.doi.org/10.9781/ijimai.2024.01.002>

[7] is an approach aimed at solving this by retrieving HUIs without using a minimum utility threshold. Instead, the user provides an input parameter k , which represents a desired number of HUIs, and the algorithm's objective is to discover the k HUIs with the largest utilities in the database. These models are more intuitive as it is easier to set k appropriately than the minimum utility threshold. However, top- k HUIM is computationally demanding compared to traditional HUIM as the minimum utility threshold is applied to prune the search space. Generally, the larger the minimum utility threshold is, the fewer candidates the algorithm has to consider. Therefore, the initial search space in top- k HUIM is equivalent to HUIM with the minimum utility threshold set to zero.

Evolutionary computation (EC) [8] is a collection of meta-heuristic models utilizing biological principles to explore search spaces efficiently. The purpose of EC is to obtain a set of approximate solutions by analyzing problems for a limited number of iterations. One such method applied to various utility mining and search problems is particle swarm optimization (PSO) [9]–[12]. Like other EC models, PSO iteratively optimizes a problem by evolving a set of candidate solutions regarding a given quality measure. New candidates are continuously created by inheriting traits from the best solutions in previous generations, which allows the algorithm to find optimal values without exploring the search space exhaustively.

This paper proposes a heuristic model based on PSO to find the top- k HUIs, called TKU-PSO. To our knowledge, this is the first work on EC in top- k HUIM. The main contributions of the paper are listed below:

- We formulate the problem of top- k HUIM from the perspective of evolutionary computation and particle swarm optimization, in which candidate quality is evaluated based on a utility fitness function.
- We introduce several new strategies to improve the general performance of heuristics in utility mining. First, to enhance the model's ability to find optimal solutions in large search spaces, the best 1-itemsets are utilized for better population initialization. Second, redundant and unnecessary particle evaluations are avoided through fitness estimation and by maintaining previously explored candidates. Finally, to reduce the algorithm's required search space, unpromising items are pruned with a threshold-raising concept called minimum solution fitness.
- We conduct a series of experiments on real- and synthetic data to evaluate the performance of the designed model against existing top- k HUIM methods. The results show that TKU-PSO outperforms the current state-of-the-art approaches in all tested datasets.

The remainder of this paper is organized as follows: Section II reviews related works. Section III presents the preliminaries and problem statement. Section IV introduces the proposed strategies and algorithm. Section V illustrates the model with an example. Section VI discusses the results of the conducted experiments. Section VII gives a conclusion of the presented work.

II. RELATED WORK

This section gives an overview of the exhaustive algorithms proposed for HUIM and top- k HUIM before reviewing the heuristic alternatives.

A. High-Utility Itemset Mining

A vital challenge in HUIM is to deal with potentially huge search spaces. A database with n distinct items contains $2^n - 1$ HUI candidates, which means naive approaches quickly struggle due to combinatorial explosion. In this respect, Liu et al. [13] provided one of the main

breakthroughs in HUIM with the Two-Phase algorithm. They introduced a technique to reduce the number of candidates based on transaction-weighted utilities (TWU). If the TWU of an itemset is less than the minimum utility threshold, then no superset extension of the itemset can be a HUI. This concept is employed during the first phase of the algorithm to only generate candidates that satisfy the TWU constraint. The second phase then identifies the actual HUIs by determining the utility of each candidate. Several other algorithms based on the two-phase approach have later been suggested, such as IHUP [14], UP-Growth [15], and MU-Growth [16]. They apply different tree structures during candidate generation to avoid creating itemsets that do not appear in the input database, thus reducing the number of necessary evaluations.

Although the two-phase algorithms establish boundaries to the search space, they often cannot reduce the number of candidates sufficiently. In addition, the models are subject to computationally expensive database scans during the evaluation phase of the candidates. In order to alleviate this, Liu and Qu [17] proposed HUI-Miner, a one-phase approach without candidate generation. They developed a utility-list data structure to hold itemset information instead of the database. The model performs two database scans to construct an initial set of utility-lists before the HUIs are identified directly through utility-list join-operations. This way, the algorithm bypasses the candidate generation phase, which requires each candidate to be cached, potentially leading to memory limitations. Moreover, utility-lists enable more efficient evaluations than database scans while providing further search space pruning through the concept of remaining utility. The approach has later been improved with algorithms that reduce the computational cost associated with join-operations, some of which are FHM [18], HUP-Miner [19], and UBP-Miner [20].

There have also been introduced one-phase approaches that avoid irrelevant itemsets, similar to the tree-based, two-phase algorithms. The d²HUP [21] algorithm enumerates itemsets as prefix extensions by using a hyper-structure database projection, which was shown to be generally more efficient than the earlier utility-list-based methods. Later, EFIM [22] reduced the cost of database scans with transaction merging and database projection techniques. The model utilizes a utility-array structure to hold item information, allowing linear time utility calculations. In addition, EFIM introduced subtree- and local utility upper bounds for further search space reduction.

B. Top-K High-Utility Itemset Mining

HUIM algorithms perform search space pruning by comparing different utility upper bounds to the user-specified minimum utility threshold. In top- k HUIM, the minimum utility threshold is initialized to zero to overcome the difficulty of selecting an appropriate value. These algorithms thus face additional search space challenges and rely on threshold-raising strategies to gradually prune unpromising candidates. However, the mining- and pruning logic are generally adopted from earlier HUIM works.

Wu et al. [7] were the first to introduce top- k HUIM with the TKU algorithm. TKU is a two-phase model that relies on five threshold-raising strategies to reduce the number of candidates with TWU pruning. The first phase of the algorithm maps potential top- k HUIs (PKHUIs) to a tree-based structure (UP-Tree) by scanning the input data twice. The second phase then determines the actual top- k HUIs by traversing the tree and evaluating the utility of the PKHUIs. To improve the performance of TKU, Ryang and Yun developed REPT [23]. REPT builds upon the same two-phase concept but applies more effective threshold-raising and thus generates fewer PKHUIs. Although the algorithm is superior to TKU, it requires an additional input parameter N , which can be challenging to select.

Due to the two-phase paradigm, TKO and REPT inherits the same limitations as their HUIM relatives. Later methods thus adopt the superior one-phase strategy. TKO [24] is a HUI-Miner extension that combines novel threshold-raising with the utility-list structure. The model reveals HUIs without producing candidates and performs pruning based on TWU and remaining utility, which alleviates the computational burden associated with the earlier two-phase algorithms.

Duong et al. [25] then introduced kHMC, which also employs the utility-list strategy. In addition, kHMC applies three threshold-raising techniques to reduce candidates and uses estimated utility co-occurrence pruning and pruning by coverage to limit the number of necessary join-operations on utility-lists. The model was compared to TKO and REPT and showed overall better efficiency.

TKEH [26] is an extension of EFIM that utilizes transaction merging and database projection techniques to reduce the cost of database scans. It employs three threshold-raising strategies and two pruning strategies to evade unpromising candidates. Moreover, the utility-list is exchanged with the utility-array structure to facilitate linear time utility calculations. The model performs particularly well in dense databases since transaction merging is effective in scenarios with many similar transactions.

To improve the discovery of extremely long patterns, Liu et al. [27] developed TONUP. TONUP is a utility list-based, opportunistic pattern growth approach that uses five strategies for maintaining shortlisted patterns. The model grows the patterns as prefix extensions, shortlists patterns with the top k utilities, and prunes the search space with novel utility upper bounds. Experiments proved the model to be significantly faster than TKU and TKO, as well as several traditional HUIM algorithms tuned with an optimal minimum utility threshold.

THUI [28] is an approach that applies a leaf itemset utility structure to maintain itemset information and a novel utility lower bound estimation method to improve the effectiveness of threshold-raising and pruning. Experiments showed the model to be one to three orders of magnitude faster than kHMC and TKO, especially on dense datasets.

Finally, top- k HUIM extensions for specialized data environments have also been suggested. E.g., PTM [29] proposed a prefix-based partitioning strategy to accommodate massive datasets, TKN [30] has been introduced for mining data with negative or positive item utilities, and TKUS [31] has been proposed for finding patterns in sequential data. However, such extensions are outside the scope of this paper.

C. Heuristic HUIM and Top-K HUIM

Although the algorithms mentioned in the previous section can discover the exact top- k HUIs, they cannot efficiently deal with huge search spaces, regardless if the approach belongs to the one-phase or two-phase paradigm. For this reason, several heuristic algorithms have been proposed to tackle the problem of HUIM, particularly evolutionary computation (EC). These methods can find optimal solutions to large search problems without exploring the entire search space, which can be crucial for swift decision-making.

Currently, TKU-CE+ [32] is the only heuristic model available for top- k HUIM. However, it does not belong to the EC domain. It is an iterative approach based on cross-entropy that generates random samples and updates parameters to produce better samples in subsequent iterations. The authors also proposed a pruning strategy based on a critical utility value (CUV). During the initialization process, the model calculates 1-itemsets utilities and sets CUV to the k -th largest utility. Unpromising candidates are then pruned based on the TWU model from traditional HUIM [13]. In addition, they used a sample refinement strategy and smoothing mutation to increase sample diversity and mining performance. The algorithm

demonstrated competitive runtimes and memory usage compared to TKU, TKO, and kHMC, although for a limited range of k . As there are no other heuristics for top- k HUIM, the rest of this section outlines the most relevant works introduced for traditional HUIM. All of these approaches utilize the basic TWU model for search space pruning.

Particle swarm optimization (PSO) is an evolutionary-based procedure extensively applied in HUIM. PSO maintains a population of particles that represent potential solutions. Each particle is assigned a fitness value and a velocity vector. The fitness determines the quality of the solution, while the velocity decides how the particle evolves. At each iteration of the algorithm, the velocity is updated based on two historical particles—the personal fittest offspring of the particle ($pBest$) and the all-time fittest particle in the entire population ($gBest$). After the new velocity is acquired, the particle is updated and evaluated, and $pBest$ and $gBest$ are redetermined. This way, the population evolves towards the optimal solution(s) by modifying particles according to the most promising candidates evaluated.

Lin et al. introduced two PSO models with HUIM-BPSO+ [10] and HUIM-BPSO- [33]. The difference between the approaches is that HUIM-BPSO+ uses an OR-NOR tree to produce valid item combinations and thus avoids evaluating irrelevant solutions. Song and Huang [34] used a similar approach in Bio-HUIF-PSO where a promising encoding vector check (PEV-check) is applied to prune the candidates that do not appear in any transaction. In addition, they improved population diversity by using roulette wheel selection to update $gBest$ among the discovered HUIs. The velocity function was also replaced with a more effective bit difference strategy. More recently, Fang et al. [35] introduced HUIM-IBPSO, which uses several adjustment strategies to escape local optima and improve the overall convergence and accuracy.

The genetic algorithm (GA) is also a biologically inspired technique in which a population of chromosomes evolves towards the optimal values using selection, crossover, and mutation operations. Kannimuthu and Premalatha [36] introduced two GA models for HUIM. Their distinction is whether a minimum utility threshold is required or not. However, both methods struggle with premature convergence to local optima. To improve this, Zhang et al. introduced HUIM-IGA [37], which employs neighborhood exploration, population diversity maintenance, individual repair, and elite strategy for better search space exploration. Another GA model was proposed with Bio-HUIF-GA [34], which uses the strategies of Bio-HUIF-PSO to avoid irrelevant candidates and boost performance.

Several other types of EC have also been proposed for HUIM. Wu et al. [38] used ant colony optimization to map the search space to a routing graph and explored it using pheromone rules. Song et al. have developed approaches with artificial bee colony algorithm [39], bat algorithm [34], and artificial fish swarm algorithm [40]. There are also heuristic HUIM techniques not based on EC, such as hill climbing and simulated annealing [41].

Altogether, the PSO-based approaches have shown the most promise for heuristic discovery of HUIs. The GA models can provide slightly higher accuracy but will generally use more time as their update procedures require additional computations. The other EC approaches tend to struggle with local optima in the iterative stage and thus miss a large portion of the available solutions. Based on this, the PSO algorithm is an opportune candidate for a heuristic top- k HUIM model.

Table I gives an overview of the current top- k HUIM algorithms and their main characteristics.

TABLE I. OVERVIEW OF TOP-K HUIM ALGORITHMS

Algorithm	Type	Base-algorithm	Year
TKU [7]	Exact (two-phase)	Up-Growth [15]	2012
REPT [23]	Exact (two-phase)	MU-Growth [16]	2015
TKO [24]	Exact (one-phase)	HUI-Miner [17]	2015
kHMC [25]	Exact (one-phase)	FHM [18]	2016
TONUP [27]	Exact (one-phase)	d2HUP [21]	2018
TKEH [26]	Exact (one-phase)	EFIM [22]	2019
THUI [28]	Exact (one-phase)	HUI-Miner [17]	2019
TKU-CE+ [32]	Heuristic	Cross-entropy [42]	2021

D. Limitations of Prior Works

Heuristics are a vital research topic in data mining as they alleviate the computational burden associated with analyzing massive datasets. However, as the last section shows, there is an abundance of heuristics available for HUIM but only one method for top-k HUIM. We also argue that all these previous works suffer the same fault—they spend too much time evaluating unpromising or redundant solutions. Fitness evaluation of a candidate can be extremely costly as the algorithm must scan the database to calculate the utility. The total number of evaluations thus significantly affects the algorithm's overall runtime. Some studies try to solve this with various termination criteria. However, due to the random nature of stochastic optimization, convergence is unpredictable and challenging to measure, and the model's accuracy will typically decline.

Another concern with current heuristics is their accuracy in large search spaces. As the search space grows, it is increasingly difficult to generate suitable initial candidates. If they share few similarities with the best solutions, the algorithm tends to fall into local optima before generating any appropriate candidates.

The goal of this paper is thus to devise a heuristic top-k HUIM model that also mitigates these limitations of previous works.

III. PRELIMINARIES AND PROBLEM STATEMENT

Let the set $I = \{i_1, i_2, \dots, i_m\}$ contain m distinct items, where i_k is a unique item such that $1 \leq k \leq m$. A transactional database $D = \{T_1, T_2, \dots, T_n\}$ is a set of n transactions, where each transaction $T_q \subseteq I$ and q is a unique transaction identifier (TID) such that $1 \leq q \leq n$. Moreover, each item $i_k \in D$ is associated with a profit value, denoted $p(i_k, D)$, and a purchase quantity for each transaction, denoted $q(i_k, T_q)$. The set $X \subseteq I$ is called an itemset and is included in transaction T_q if $X \subseteq T_q$. In addition, an itemset with p items is called a p -itemset.

The database shown in Table II is used as a running example in this paper. It contains six transactions and six distinct items named from A to F , with the corresponding purchase quantities inside the parentheses. Table III shows the associated profit value of each item.

TABLE II. A QUANTITATIVE TRANSACTIONAL DATABASE

TID	Trans (item : quantity)	tu
T_1	(D:2), (E:3)	16
T_2	(A:1), (D:2), (E:2)	17
T_3	(A:1), (B:2), (F:1)	6
T_4	(C:4), (E:3)	14
T_5	(B:3), (C:1), (D:1)	10
T_6	(F:9)	9

TABLE III. PROFIT TABLE

Item	A	B	C	D	E	F
Unit profit	3	1	2	5	2	1

Definition 1. The utility of an item i_k in a transaction T_q is denoted $u(i_k, T_q)$ and is calculated by (1).

$$u(i_k, T_q) = q(i_k, T_q) \times p(i_k, D) \quad (1)$$

Example 1. The utility of item D in transaction T_1 is calculated as $2 \times 5 = 10$.

Definition 2. The utility of an itemset X in a transaction T_q is denoted $u(X, T_q)$ and is calculated by (2).

$$u(X, T_q) = \sum_{i_k \in X, X \subseteq T_q} u(i_k, T_q) \quad (2)$$

Example 2. The utility of itemset (BC) in transaction T_5 is calculated as $3 \times 1 + 1 \times 2 = 5$.

Definition 3. The utility of an itemset X in a database D is denoted $u(X)$ and is calculated by (3).

$$u(X) = \sum_{X \subseteq T_q, T_q \in D} u(X, T_q) \quad (3)$$

Example 3. The utility of itemset (DE) is calculated as $2 \times 5 + 3 \times 2 + 2 \times 5 + 2 \times 2 = 30$.

Definition 4. The TID-set of an itemset X in a database D is denoted $TID(X)$ and is calculated by (4).

$$TID(X) = \{q \mid q \geq 1, q \leq n, X \subseteq T_q, T_q \in D\} \quad (4)$$

Example 4. The TID-set of itemset (D) is $\{1, 2, 5\}$, as (D) occurs in T_1, T_2 and T_5 .

Definition 5. The support count of an itemset X is denoted $sup(X)$ and is calculated by (5).

$$sup(X) = |TID(X)| \quad (5)$$

Example 5. The support of itemset (D) is calculated as $|\{1, 2, 5\}| = 3$.

Definition 6. The transaction utility of a transaction T_q is denoted $tu(T_q)$ and is calculated by (6).

$$tu(T_q) = \sum_{i_k \in T_q} u(i_k, T_q) \quad (6)$$

Example 6. The transaction utility of T_5 is calculated as $3 \times 1 + 1 \times 2 + 1 \times 5 = 10$.

Definition 7. The transaction-weighted utility (TWU) of an itemset X is denoted $TWU(X)$ and is calculated by (7).

$$TWU(X) = \sum_{q \in TID(X)} tu(T_q) \quad (7)$$

Example 7. The TWU of itemset (E) is calculated as $16 + 17 + 14 = 47$.

Definition 8. Given a minimum utility threshold δ , an itemset X is a high transaction-weighted utilization itemset (HTWUI) if $TWU(X) \geq \delta$; otherwise, X is a low transaction-weighted utilization itemset (LTWUI). In addition, a HTWUI/LTWUI with p items is denoted p -HTWUI/ p -LTWUI.

Example 8. If the minimum utility threshold is set to 20, then itemset (B) is a 1-LTWUI since $TWU(B) = 16$, while itemset (A) is a 1-HTWUI as $TWU(A) = 23$.

Definition 9. Given an minimum utility threshold δ , an itemset X is a high-utility itemset (HUI) if $u(X) \geq \delta$.

Example 9. If the minimum utility threshold is 20, then itemset (D) is a HUI as $u(D) = 25$.

Definition 10. An itemset X is a top- k HUI in a database D if its utility is among the k largest in D .

Example 10. If k is 3, then the set of top- k HUIs is $\{(DE:30), (D:25), (ADE:17)\}$.

Problem statement: Given a desired number of HUIs (k) and a database D , the problem of top- k HUIM is to determine the k HUIs with the largest utilities in D .

IV. PROPOSED ALGORITHM FOR TOP-K HUIM

The proposed TKU-PSO is an iterative approach that prunes the search space before a population of particles is generated based on the remaining candidates. The top- k HUIs are discovered by evaluating and updating the population for a desired number of iterations. We will explain the model in five parts, where the first four describe the main developed strategies, and the last section introduces the complete model.

A. Minimum Solution Fitness

To maintain the discovered top- k HUIs, we employ a set with the maximum capacity of k (the desired number of HUIs), where each solution is sorted in descending order of utility. In other words, the solution with the smallest utility is always at the tail of the set. For simplicity throughout the paper, we call the utility of the tail-itemset the minimum solution fitness. It is defined as follows:

Definition 11. The minimum solution fitness is denoted $MSF(H)$ and is calculated by (8).

$$MSF(H) = \begin{cases} u(H_k), & \text{if } |H| = k \\ 0, & \text{otherwise,} \end{cases} \quad (8)$$

where H is the set of current top- k HUIs sorted in descending order of utility, and k is the desired number of HUIs

The minimum solution fitness is zero until the top- k set reaches its capacity, and the model only stores a new solution if its utility exceeds the current value. Once the set is full, new solutions replace the tail-itemset. This way, the minimum solution fitness is a dynamic threshold that grows as the algorithm progresses. The following sections explain how the model utilizes the minimum solution fitness to avoid fitness evaluations and prune candidates.

B. Population Initialization Strategy

The designed model represents each particle with a bit vector, called an encoding vector. The encoding vector length corresponds to the number of 1-HTWUI in the database, and each bit describes a specific item. If position i of an encoding vector is 1, then item i is included in the particle; otherwise, item i is not included. For example, assuming all items in Table II are 1-HTWUI, the encoding vector of itemset (ABF) is {1, 1, 0, 0, 0, 1}.

As there is no minimum utility threshold in top- k HUIM, all items are initially 1-HTWUI. However, the proposed model removes 1-LTWUIs by setting the minimum utility threshold to the critical utility value (CUV) [32]. CUV is found by calculating all 1-itemset utilities and sorting them in descending order of utility. We utilize this to initialize the first particles to the 1-itemsets with the largest utilities in the database. Previous algorithms initialize the first candidates to random itemset sizes between 1 and the number of 1-HTWUIs, which means they will generate huge itemsets in databases with many 1-HTWUIs. As a result, the model likely converges to local optima as the best solutions generally are much smaller than the number of 1-HTWUIs. Initialization with 1-itemsets can thus provide particles more similar to the relevant solutions and simplify the evolutionary process. In addition, the algorithm's performance becomes more consistent as the first population is selected deterministically rather than stochastically.

However, if the population size is larger than the number of 1-HTWUIs, not all particles can be initialized to a unique 1-itemset. In this scenario, we generate the leftover particles with roulette

wheel selection. Moreover, any particle generated with roulette wheel selection is PEV-checked. The PEV-check ensures the particle appears in at least one transaction, and the algorithm avoids evaluating irrelevant solutions. The implementation details of roulette wheel selection and PEV-check are described by Song and Huang [34].

Algorithm 1 shows the population initialization procedure. First, the database is scanned once to calculate the utility and TWU of each 1-itemset (line 1). The minimum utility threshold is then set to the k -th largest utility, and each 1-LTWUI is pruned from the database (line 2). The 1-HTWUIs are then sorted in descending order of utility before the population, $pBest$, and solutions are initialized to empty (lines 3 and 4). Thereafter, the main loop of the procedure starts, where pop_size particles are generated (lines 5-18). At each iteration, it is checked whether the set of 1-HTWUI is empty (line 6). If not, the first 1-HTWUI in I is popped, and the particle is initialized to the 1-itemset representing this 1-HTWUI. (lines 7 and 8). Otherwise, the particle is generated with roulette wheel selection and PEV-checked (lines 9-12). Next, the created particle is evaluated by calculating its fitness (line 13). If the fitness is larger than the minimum solution fitness, the particle is put in the set of top- k HUIs as described in Section A (lines 14-16). Finally, the particle is placed in the population and its corresponding $pBest$ before the next iteration starts (line 17). After the entire population is created, the set of top- k HUIs is filled with the remaining 1-itemsets until it is full, or there are no more 1-itemsets (lines 19-21). This step is performed to increase the minimum solution fitness quickly. Finally, the population, $pBest$, and current top- k HUIs are returned, and the procedure terminates (line 22).

Algorithm 1. Population initialization, *init()*

Input: D : a transactional database, pop_size : the population size,
 k : the number of desired HUIs

Output: Pop : the first population, $pBest$: initial offspring,
 H : the current top- k HUIs

```

1: calculate utility and TWU of each item in  $D$ ;
2: remove items with TWU less than  $k$ th largest utility;
3:  $I \leftarrow$  each 1-HTWUI, in descending order of utility;
4:  $Pop, pBest, H \leftarrow \emptyset$ ;
5: for  $i = 1$  to  $pop\_size$  do
6:   if  $|I| > 0$  then
7:      $p_i \leftarrow$  generate to the first item in  $I$ ;
8:     remove the first item in  $I$ ;
9:   else
10:     $p_i \leftarrow$  generate with roulette wheel selection;
11:     $p_i \leftarrow$  PEV-check  $p_i$ ;
12:   end if
13:    $fit \leftarrow$  calculate fitness of  $p_i$  using Eq. (9);
14:   if  $fit > MSF(H)$  then
15:     insert  $p_i$  into  $H$ ;
16:   end if
17:    $Pop, pBest, \leftarrow p_i$ ;
18: end for
19: if  $pop\_size < k$  and  $|I| > 0$  then
20:   fill  $H$  with the remaining 1-itemsets in  $I$ ;
21: end if
22: Return  $Pop, pBest, H$ ;
```

C. Fitness Evaluation Strategies

The model evaluates the quality of each particle in the population with a fitness function.

Definition 12. The fitness of a particle p_i is denoted $fit(p_i)$ and is defined in (9).

$$fit(p_i) = u(X) \quad (9)$$

where X is the itemset in the encoding vector of p_i .

Calculating the utility of an itemset is a costly operation in heuristic utility mining algorithms. The time complexity is approximately $O(s \times a)$, where s is the support of the itemset, and a is the average transaction length in the database. Therefore, it is desirable to skip the evaluation of certain unpromising candidates to improve the execution time of the model. First, many redundant particles are created during the runtime, especially if the algorithm converges. As it is unnecessary to assess these solutions repeatedly, the proposed model maintains each created particle in a hash set. If the set contains a specific particle, the solution is redundant, and the algorithm does not perform the fitness evaluation. By doing this, the model quickly terminates when it converges as it will primarily create explored solutions.

To further reduce the number of evaluations, we employ a strategy to approximate the fitness, which we call fitness estimation.

Definition 13. The maximum utility of an item i in a database D is denoted $mu(i)$ and is calculated by (10).

$$mu(i) = \max\{u(i, T_q)\}_{\forall T_q \in D} \quad (10)$$

Example 11. The maximum utility of item D in Table II is calculated as $\max\{10, 10, 5\} = 10$

Definition 14. The average utility of an item i in a database D is denoted $au(i)$ and is calculated by (11).

$$au(i) = \left\lfloor \frac{\sum_{T_q \in D} u(i, T_q)}{sup(i)} \right\rfloor \quad (11)$$

Example 12. The average utility of item D in Table II is $\left\lfloor \frac{10+10+5}{3} \right\rfloor = 9$.

Definition 15. The estimated utility of an itemset X is denoted $Est(X)$ and is calculated by (12).

$$Est(X) = sup(X) \times \left(\sum_{i_k \in X} au(i_k) + \sigma \right) \quad (12)$$

where the deviation σ is defined by (13).

$$\sigma = \left\lfloor \frac{\sum_{i_k \in 1-HTWUI} mu(i_k) - au(i_k)}{|1-HTWUI|} \right\rfloor \quad (13)$$

Example 13. Assuming all items in Table II are 1-HTWUI, the deviation is calculated as $\left\lfloor \frac{(8-3)+(3-3)+(8-5)+(10-9)+(6-6)+(9-5)}{6} \right\rfloor = 2$. Thus, the estimated utility of itemset (D) is calculated as $3 \times (9 + 2) = 33$.

The model uses the estimated utility to determine whether evaluating a particular particle is worthwhile. It does this by comparing the estimate to the fitness of $pBest$ and the minimum solution fitness. If the estimate is less than both values, the particle will likely not improve the population or be a top- k HUI, and the evaluation is thus skipped. Based on Example 13, the model ignores the evaluation of itemset (D) if the fitness of $pBest$ and the minimum solution fitness is at least 33.

The purpose of the deviation is to avoid underestimates. An underestimate occurs when an estimate is less than the particle's actual fitness. Otherwise, the estimate is an overestimate. The model keeps track of the number of over- and underestimates during runtime and occasionally updates the deviation according to (14).

$$\sigma = \begin{cases} \frac{\sigma}{2}, & \text{if } \sigma > 1 \text{ and } \frac{u}{o} < 0.01 \\ \sigma, & \text{otherwise,} \end{cases} \quad (14)$$

where the number of over- and underestimates are denoted as o and u , respectively.

Example 14. Assuming $u = 0$ and $o = 100$. The deviation of Table I is updated as $\frac{2}{2} = 1$, and the estimated utility of itemset (D) is calculated as $3 \times (9 + 1) = 30$.

This way, the model adapts to the data and produces more accurate estimates as the deviation is progressively tuned. Each fitness estimate is calculated in linear time on the size of the itemset, which is negligible compared to the complexity of finding the actual utility. The algorithm can thus save significant time when generating many low-fitness particles.

D. Particle Update Strategy

The designed model updates each particle towards $pBest$ and $gBest$ using the concept of bit difference [34]. It is defined as follows:

Definition 16. The bit difference of two particles p_i and p_j , denoted $BitDiff(p_i, p_j)$, is defined as the bitwise-XOR operation on the encoding vectors of the particles.

Example 15. Let $p_1 = \{0, 1, 1, 0\}$ and $p_2 = \{1, 0, 1, 0\}$, then $BitDiff(p_1, p_2) = \{1, 1, 0, 0\}$.

In other words, bit difference creates a bit vector of non-identical bits between two particles. The update procedure uses bit difference to compare a particle to $pBest$ and $gBest$, and the bits set to 1 in the vector represent the items that can change in the particle.

However, if the population only evolves based on the previously best solutions, the model typically falls in a local optimum due to insufficient diversity. We increase the amount of exploration by performing a random modification to the particle after the update towards $pBest$ and $gBest$ is complete. The model only executes this step if the current particle is a redundant solution. Thus, we avoid randomly altering new solutions to previously explored ones. The total number of bits b_i to change in a particle p_i is determined by (15).

$$b_i = b_{i1} + b_{i2} + b_{i3} \quad (15)$$

where b_{i1} , b_{i2} , and b_{i3} are defined in (16), (17), and (18), respectively.

$$b_{i1} = \lfloor r_1 \times \sum BitDiff(p_i, pBest_i) \rfloor \quad (16)$$

$$b_{i2} = \lfloor r_2 \times \sum BitDiff(p_i, gBest) \rfloor \quad (17)$$

$$b_{i3} = \begin{cases} 1, & \text{if } p_i \in E \\ 0, & \text{otherwise} \end{cases} \quad (18)$$

where r_1 and r_2 are random numbers between $[0,1]$, and E is the hash set of explored particles. Note that b_{i3} is determined after $b_{i1} + b_{i2}$ changes are made to the particle.

The update procedure selects b_i items and flips their corresponding bit in the particle's encoding vector. However, some 1-HTWUIs can have a TWU value less than the minimum solution fitness as it grows during runtime. An itemset containing any such 1-HTWUI cannot be part of a top- k HUI. Therefore, the algorithm always performs the bit clear operation on these items in the particle. Doing this lowers the number of potential candidates and thus improves the algorithm's ability to generate the actual solutions.

Algorithm 2 shows the particle update procedure. First, b_{i1} of the different items between p_i and $pBest_i$ are randomly selected and put into the set I (lines 1 and 2). Each item in I is flipped or cleared in the particle, depending on the item's TWU value and the current minimum solution fitness (lines 3-5). Next, the above process repeats for p_i and $gBest$ (lines 6-10), before b_{i3} is calculated by identifying whether the

current particle is redundant (line 11). If it is redundant, one additional random item is flipped or cleared in the particle (line 13). Finally, the updated particle is PEV-checked and returned (lines 14 and 15).

Algorithm 2. Particle update, *update()*

Input: p_i : the particle

Output: p'_i : the updated particle

```

1:  $b \leftarrow$  calculate  $b_{i1}$  using Eq. (16);
2:  $I \leftarrow b$  random items set to 1 in  $BitDiff(p_i, pBest_i)$ ;
3: for each  $item \in I$  do
4:    $p_i \leftarrow$  flip or clear  $item$  in  $p_i$ ;
5: end for
6:  $b \leftarrow$  calculate  $b_{i2}$  using Eq. (17);
7:  $I \leftarrow b$  random items set to 1 in  $BitDiff(p_i, gBest_i)$ ;
8: for each  $item \in I$  do
9:    $p_i \leftarrow$  flip or clear  $item$  in  $p_i$ ;
10: end for
11:  $b \leftarrow$  calculate  $b_{i3}$  using Eq. (18);
12:  $item \leftarrow b$  random 1-HTWUI;
13:  $p_i \leftarrow$  flip or clear  $item$  in  $p_i$ ;
14:  $p_i \leftarrow$  PEV-check  $p_i$ ;
15: return  $p'_i$ ;

```

E. TKU-PSO

Algorithm 3 shows the designed TKU-PSO in its entirety. The model takes as input a transactional database, the number of desired

Algorithm 3 Proposed TKU-PSO Algorithm

Input: D : a transactional database, k : the desired number of HUIs,
 pop_size : the population size, $iter$: the number of iterations.

Output: H : set of top- k HUIs

```

1:  $Pop, pBest, H \leftarrow \text{init}(D, pop\_size, k)$ ;
2:  $gBest \leftarrow$  the fittest particle in  $Pop$ ;
3:  $E \leftarrow Pop$ ;
4:  $\sigma \leftarrow \text{calc. using Eq. (13)}$ ;
5: for  $i = 1$  to  $iter$  do
6:   for  $j = 1$  to  $pop\_size$  do
7:      $Pop_j \leftarrow \text{update}(Pop_j)$ ;
8:     if  $Pop_j \notin E$  then
9:        $X \leftarrow$  the itemset in  $Pop_j$ ;
10:       $est \leftarrow$  estimate the utility of  $X$  using Eq. (12);
11:      if  $est > MSF(H)$  or  $est > fit(pBest_j)$  then
12:         $fit \leftarrow \text{calc. fitness of } Pop_j \text{ using Eq. (9)}$ ;
13:        if  $fit > MSF(H)$  then
14:          insert  $Pop_j$  into  $H$ ;
15:        end if
16:         $pBest_j \leftarrow$  fittest of  $Pop_j$  and  $pBest_j$ ;
17:         $gBest \leftarrow$  fittest of  $Pop_j$  and  $gBest$ ;
18:      end if
19:       $E \leftarrow E \cup Pop_j$ ;
20:    end if
21:  end for
22:  $gBest \leftarrow$  update with roulette wheel selection;
23:  $\sigma \leftarrow$  update using Eq. (14);
24: end for
25: return  $H$ ;

```

HUIs, the population size, and the number of iterations. First, the population, $pBest$, and the set of top- k HUIs are initialized by calling the initialization procedure of Algorithm 1 (line 1). Next, $gBest$ is set to the fittest particle, and the set of explored particles is filled with the current population (lines 2 and 3). The deviation of the maximum- and average utilities are then calculated (line 4) before the main loop of the procedure starts, where the population is iteratively updated and evaluated (lines -24). At each iteration, the particles are updated using Algorithm 2 (line 7). If a new particle is redundant, it is not evaluated further, and the procedure continues with the next particle in the population (line 8). Otherwise, the particle's fitness is estimated to determine if evaluation should proceed (lines 9 and 10). The particle's exact fitness is only found if the estimate is greater than the fitness of $pBest$ or the current minimum solution fitness (lines 11 and 12). If the fitness is greater than the minimum solution fitness, the particle is a new top- k HUI and is inserted into the solution set as described in Section A (lines 13-15). Then, $pBest$ and $gBest$ are updated accordingly (lines 16 and 17), and the particle is marked as explored (line 19). When the entire population is updated and evaluated, $gBest$ is reselected to one of the current top- k HUIs using roulette wheel selection (line 22). This step is not performed if $gBest$ was updated naturally during the current iteration. The deviation is then updated according to the number of over- and underestimates before the next iteration starts (line 23). Finally, when all iterations are complete, the set of top- k HUIs is returned, and the algorithm terminates (line 25).

V. AN ILLUSTRATED EXAMPLE

This section demonstrates the process of the designed model on the database in Table II. The population size and k (the number of desired HUIs) are 3 and 2, respectively.

First, we find the TWU $\{A:23, B:16, C:24, D:43, E:47, F:15\}$ and utility $\{A:6, B:5, C:10, D:25, E:16, F:10\}$ of each 1-itemset. The minimum utility threshold is then set to the k -th largest utility, which is 16. Based on this, item F is pruned from the database since its TWU is less than the minimum utility threshold. The set of 1-HTWUIs is thus $\{A, B, C, D, E\}$. As the population size is less than the number of 1-HTWUIs, each particle is initialized to the 1-itemsets with the greatest utilities. Table IV shows the initial population.

TABLE IV. THE INITIAL PARTICLES IN THE POPULATION

Particle	A	B	C	D	E
P_1	0	0	0	1	0
P_2	0	0	0	0	1
P_3	0	0	1	0	0

The fittest particles are placed in the set of top- k HUIs $\{D:25, E:16\}$, and the minimum solution fitness changes to the tail-itemset's utility (16). Next, $pBest$ is initialized as a copy of the population and $gBest$ is set to P_1 . Before the update procedure starts, each current particle is marked as explored.

The update of P_2 with $r_1 = 0.7$ and $r_2 = 0.5$ goes as follows: First, $BitDiff(P_2, pBest_2)$ is calculated to $\{0,0,0,0\}$ and $b_{21} = [0.7 \times 0]$, which is 0. Therefore, no items change in P_2 . Next, $BitDiff(P_2, gBest)$ is calculated to $\{0,0,0,1,1\}$ and $b_{22} = [0.5 \times 2]$, which is 1. As a result, one non-identical bit between P_2 and $gBest$ must change, either the bit representing item D or E . Assuming item D is selected, its bit is flipped because the TWU of D (43) is larger than the minimum solution fitness (16), and P_2 becomes $\{0,0,0,1,1\}$. P_2 is not a redundant solution, and b_{23} is thus 0. The update is then complete as this encoding vector is a PEV.

Suppose the updated population is $\{P_1: \{0,0,0,0,1\}, P_2: \{0,0,0,1,1\}, P_3: \{0,1,1,0,0\}\}$. Consequently, P_1 is not evaluated because it was explored in the last population. The maximum utilities of the

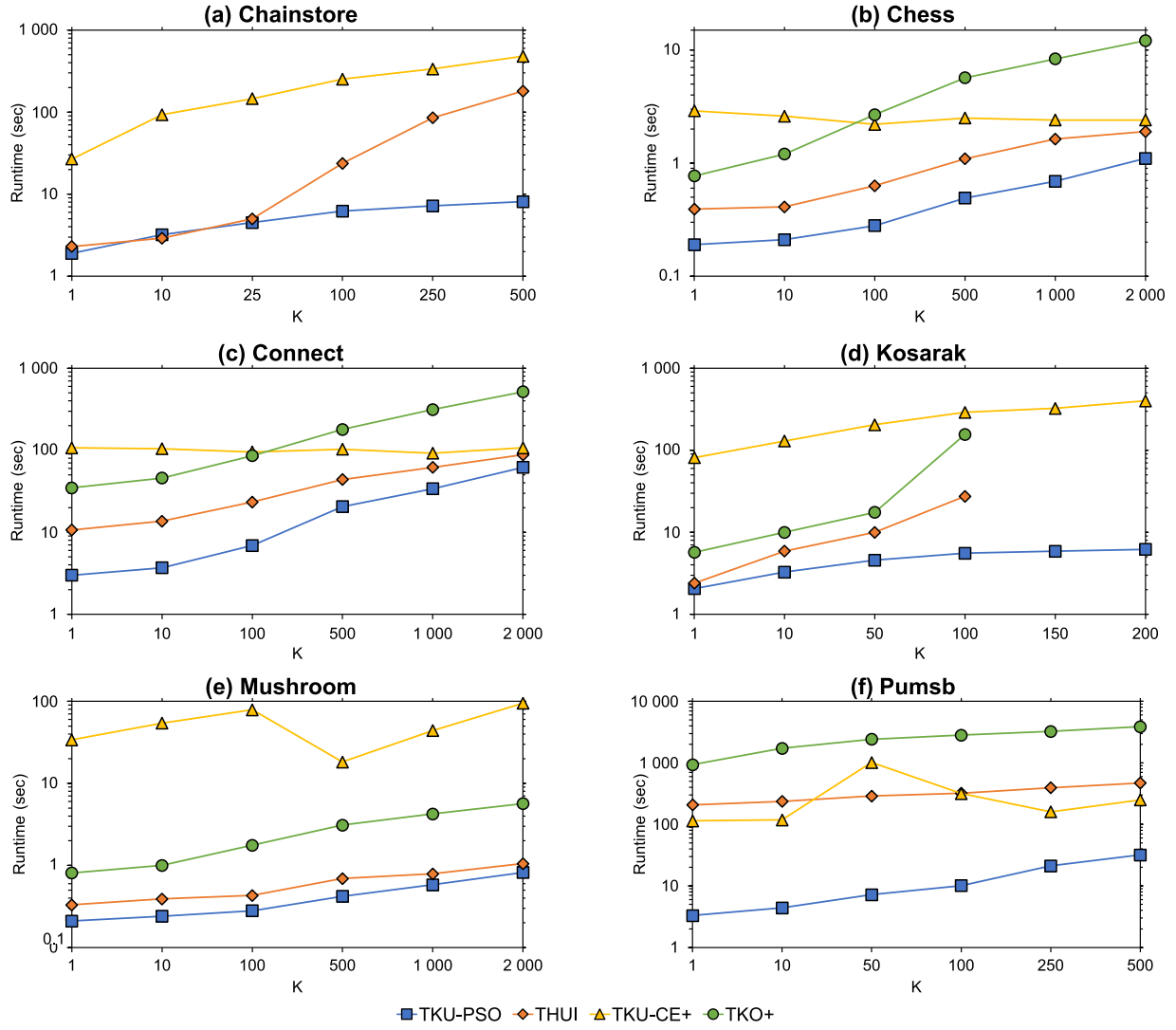


Fig. 1. The runtimes of the compared algorithms.

1-HTWUI are $\{A:3, B:3, C:8, D:10, E:6\}$, the average utilities are $\{A:3, B:3, C:5, D:9, E:6\}$, and the deviation is 1. As a result, the estimated fitness of P_2 and P_3 is 34 and 10, respectively. As the estimate of P_3 does not exceed the minimum solution fitness (16) or the fitness of $pBest_3$ (10), its fitness evaluation is skipped. The fitness of P_2 is 30, which is greater than the minimum solution fitness. The top- k HUIs are thus updated to $\{DE:30, D:25\}$, and the new minimum solution fitness is 25. In addition, $pBest_2$ and $gBest$ change to P_2 . At last, the population is put in the set of explored particles, and the next iteration begins. After the algorithm terminates, the discovered top- k HUIs are (DE) and (D) .

VI. EXPERIMENTAL RESULTS

This section evaluates the performance of the designed TKU-PSO against THUI, TKO, and TKU-CE+. The authors of TKO provided a significantly improved version of the basic TKO algorithm. We call this version TKO+ throughout the experiments. The source code of THUI was sent to us by the author while we downloaded TKU-CE+ from the SPMF data mining library [43]. The source code for TKU-PSO is available at GitHub¹. All the compared algorithms are written in Java and were executed with a heap size of 2 GB on JDK 17.0.1. We performed the experiments on a 64-bit Windows 10 computer with a Ryzen 5 5600x CPU and 16 GB of 3200 MHz CL 16 RAM.

¹ <https://github.com/Simencar/TKU-PSO>

Table V shows the characteristics of the datasets used in the comparisons. They are a mixture of real and synthetic data downloaded from SPMF. We have categorized each database as dense or sparse based on the ratio of the average transaction length to the number of distinct items in the database. Generally, sparse databases have more diverse transactions.

TABLE V. DATABASE CHARACTERISTICS

Dataset	#Items	#Trans	Avg.Trans.Len.	Type
Chainstore	46,086	1,112,949	7.23	Sparse
Chess	75	3,196	37	Dense
Connect	129	67,557	43	Dense
Kosarak	41,270	990,002	8.1	Sparse
Mushroom	119	8,416	23	Dense
Pumsb	2,113	49,046	74	Sparse

In all the tests, the proposed model is set to 10,000 iterations with a population size of 20. The iterations and sample size in TKU-CE+ are 2,000, and the quantile parameter is 0.2, as suggested by the authors. We used a lower iteration number for TKU-CE+ because it is unclear how the sample size compares to the population size of TKU-PSO. Only a proportion of the total samples are updated each iteration. In addition, TKU-CE+ uses a termination criterion that stops the execution prematurely if it determines it has converged, and the

algorithm rarely completes all iterations. Our model always performs the specified 10,000 iterations. For these reasons, the tested input parameters are fair.

A. Runtime

First, we compare the runtimes of the algorithms on the six datasets with various values of k . Fig. 1 shows the results.

Fig. 1(a) displays the comparison for Chainstore, where TKU-PSO and THUI used a similar amount of time for small values of k , but our model was up to 22 times faster as k increased. The heuristic TKU-CE+ was up to 59 times slower than TKU-PSO in Chainstore. It also terminated in less than 20 iterations on all tests. TKO+ is not included in Fig. 1(a) as it ran out of memory.

The results in the dense databases Chess and Connect are almost identical to each other, Fig. 1(b-c). Our model was the fastest for all values of k , followed by THUI. Then, TKO+ was quicker than TKU-CE+ when k was less than 100, while they swapped places for higher numbers of HUIs.

Fig. 1(d) demonstrates a clear advantage of the heuristic models in Kosarak. When k was 150 and 200, THUI and TKO+ could not finish due to the search space size. We ran THUI for over 14 hours without getting a result, while TKO+ was stopped after 3 hours. Although TKU-CE+ could complete the tests on Kosarak, it repeatedly terminated after the first iteration and was still up to 64 times slower than TKU-PSO. In addition, our model outperformed THUI and TKO+ for smaller values of k .

The Mushroom dataset also shows that TKU-PSO was the most efficient model, closely followed by THUI, Fig. 1(e). TKU-CE+ was at worst 282 times slower than TKU-PSO, while the runtime also fluctuated due to the unpredictability of the termination criterion.

Finally, Fig. 1(e) shows that TKU-PSO was much faster than the other approaches in Pumsb. THUI, TKU-CE+, and TKO+ were up to 63, 141, and 390 times slower, respectively. This was the only dataset where TKU-CE+ could finish quicker than THUI, but the runtime was inconsistent, like on Mushroom.

Overall, our model achieved the best results in terms of runtime. TKU-CE+ is slower in all tests while also performing fewer iterations. THUI is generally the closest to our model, but it cannot deal with colossal search spaces, as seen on Kosarak. Kosarak has many candidates with similar utility, and the threshold-raising pruning of THUI thus becomes ineffective. The main contributions to the speed of TKU-PSO are the strategies for redundant particles and fitness estimation, which reduces the number of necessary particle evaluations. The dynamic minimum solution fitness can also improve the runtime of the model. During particle update, we avoid 1-HTWUIs with TWU less than the minimum solution fitness. Thus, the algorithm converges quicker to the point where it creates primarily redundant solutions, which are not evaluated.

B. Accuracy

The heuristic models cannot guarantee the discovery of the correct patterns before termination. Therefore, some of the found itemsets may not correspond with the actual top- k HUIs in the database. This section compares the percentage of correct top- k HUIs between TKU-PSO and TKU-CE+. In addition, we test the proposed model without the new population initialization strategy. This model is called TKU-PSO- and uses the traditional roulette wheel selection approach. We obtained the accuracy by comparing the results of the heuristic algorithms with the output of THUI. On Kosarak, the exact patterns were retrieved with the threshold-based EFIM [22] as THUI and TKO+ could not finish for large k . The accuracy was measured with the following formula:

$$Accuracy = \frac{c}{k} \times 100 \quad (19)$$

where c is the number of correct top- k HUIs discovered by the heuristic algorithm, and k is the desired number of HUIs.

Table VI shows that the proposed TKU-PSO found significantly more correct top- k HUIs than TKU-CE+. In Kosarak, Mushroom, and Pumsb, the accuracy of our model was always 100%, while TKU-CE+ missed nearly all relevant patterns. In Chess and Connect, TKU-CE+ found the actual top- k HUIs for k up to 10, but the accuracy gradually fell to 22.5% and 20.1% as k increased. In contrast, TKU-PSO returned one incorrect itemset when k was 2,000 and maintained 100% accuracy in the other tests. In Chainstore, the proposed model performed slightly worse than in the other databases but still provided an accuracy of 96% or more. TKU-CE+ found the correct HUI at the smallest k but missed all relevant itemsets for k above 25.

TABLE VI. THE ACCURACY OF TKU-PSO, TKU-PSO- AND TKU-CE+ COMPARED

Chainstore						
k	1	10	25	100	250	500
TKU-PSO	100 %	100 %	100 %	99 %	98 %	96 %
TKU-CE+	100 %	50 %	24 %	0 %	0 %	0 %
TKU-PSO-	100 %	100 %	100 %	98 %	93.6 %	88.8 %
Chess						
k	1	10	100	500	1,000	2,000
TKU-PSO	100 %	100 %	100 %	100 %	100 %	99.9 %
TKU-CE+	100 %	100 %	90 %	51.6 %	34 %	22.5 %
TKU-PSO-	100 %	100 %	100 %	100 %	100 %	99.9 %
Connect						
k	1	10	100	500	1,000	2,000
TKU-PSO	100 %	100 %	100 %	100 %	100 %	99.9 %
TKU-CE+	100 %	100 %	80 %	39.8 %	30 %	20.1 %
TKU-PSO-	100 %	100 %	100 %	100 %	100 %	99.9 %
Kosarak						
k	1	10	50	100	150	200
TKU-PSO	100 %	100 %	100 %	100 %	100 %	100 %
TKU-CE+	100 %	0 %	0 %	0 %	0 %	0 %
TKU-PSO-	100 %	0 %	0 %	0 %	0 %	0 %
Mushroom						
k	1	10	100	500	1,000	2,000
TKU-PSO	100 %	100 %	100 %	100 %	100 %	100 %
TKU-CE+	0 %	0 %	0 %	0 %	0.01 %	0.01 %
TKU-PSO-	100 %	100 %	100 %	100 %	100 %	100 %
Pumsb						
k	1	10	50	100	250	500
TKU-PSO	100 %	100 %	100 %	100 %	100 %	100 %
TKU-CE+	0 %	0 %	0 %	0 %	0 %	0 %
TKU-PSO-	100 %	100 %	0 %	0 %	0 %	0 %

Altogether, TKU-PSO and TKU-CE+ discovered 13,113 and 2,165 correct top- k HUIs, respectively, corresponding to an overall accuracy of 99.8% and 16.5%. In other words, our model outperforms TKU-CE+ by a wide margin in these experiments. TKU-PSO can consistently find the relevant itemsets even if the search space is huge. The Kosarak results demonstrate this as the correct solutions were returned within 10 seconds, while the non-heuristic algorithms were unable to finish in any reasonable amount of time, Fig. 1(d). The main contributor

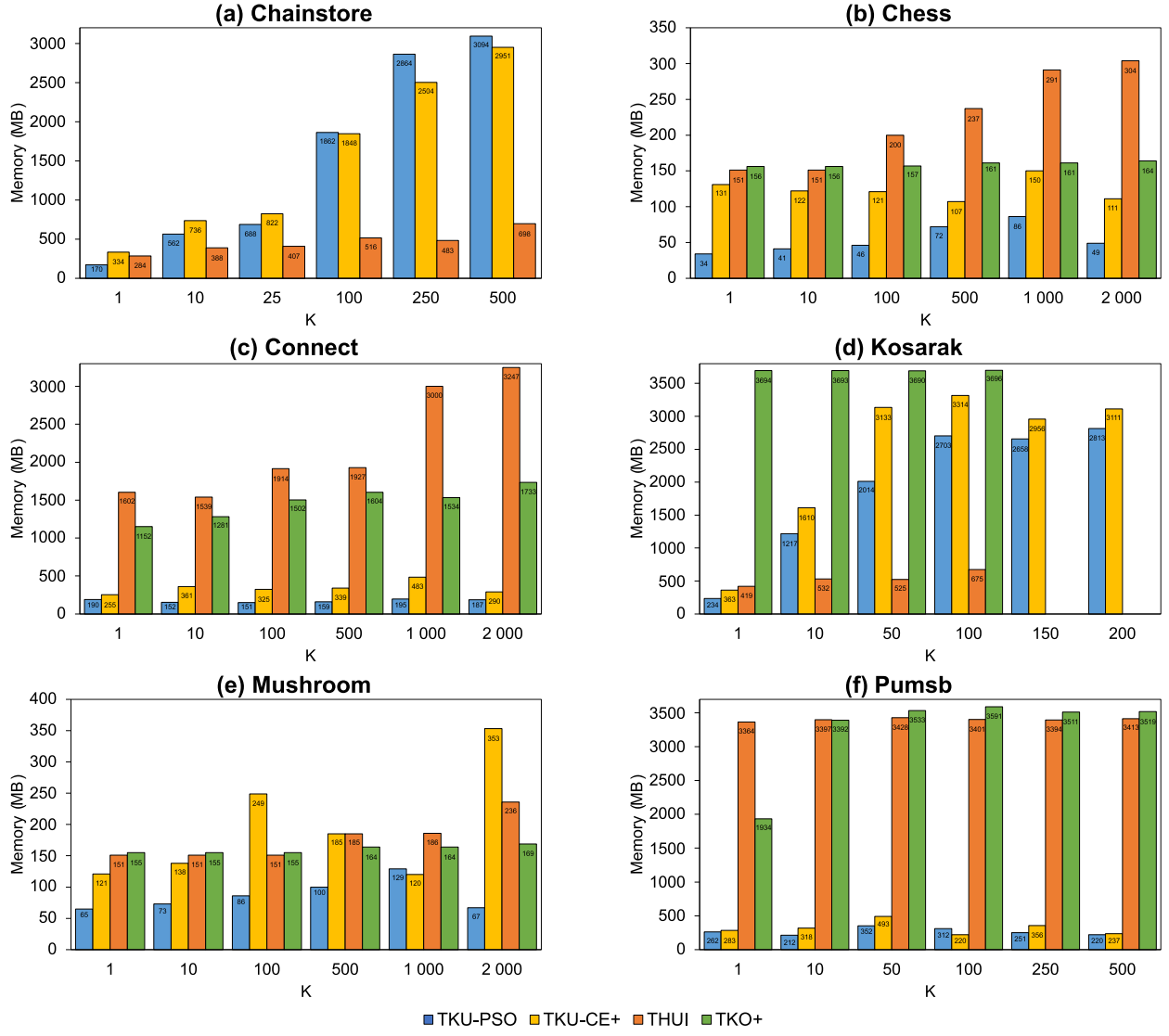


Fig. 2. The memory usage of the compared algorithms.

to this is the proposed population initialization strategy. TKU-PSO has better accuracy than TKU-PSO- in all the sparse databases. These databases have massive numbers of 1-HTWUIs when k is large, but their best itemsets are relatively small in comparison. Therefore, it is advantageous to avoid initialization with roulette wheel selection as it will create too big particles and lead the model to a local optimum. TKU-PSO- discovered most of the correct solutions on Chainstore due to the PEV-check reducing the particle sizes. Nonetheless, the new population initialization strategy always provided higher or identical accuracy.

C. Memory

Finally, we compare the maximum memory usage of each algorithm on the same datasets and k as in the previous experiments. THUI and TKO+ are missing from some graphs for the reasons stated in Section A. The memory was measured using the native Java Runtime class.

According to the results in Fig. 2, TKU-PSO used the least memory on Chess, Connect, Mushroom and Pumsb, while THUI was most efficient on Chainstore and Kosarak. This is primarily caused by the database size and the algorithm's strategy for holding item information. The heuristic models store the pruned database on the heap while THUI and TKO+ construct utility-list variations. Generally, the utility-list approach is more efficient when the database is sparse and large,

as seen on the highest k in Fig. 2(a)(d). However, our model used less memory for the smallest k in Chainstore and Kosarak because pruning reduced the database size considerably. As k increases, pruning is less effective, and memory requirements grow. TKO+ does not perform the initial pruning used by the other models. For this reason, it ran out of memory in Chainstore and performed the worst in Kosarak.

Comparing the heuristic models, TKU-PSO used overall less memory than TKU-CE+ in Fig. 2(b-f). On Chainstore, our model generates a high number of unique candidates due to the size of the search space. The memory usage then increases as the algorithm stores all explored particles. This does not happen to the same extent on the similar-sized Kosarak as the model converges early, and overall fewer candidates are examined.

Altogether, TKU-PSO was the most memory-efficient algorithm. The utility-list of THUI could use less memory in extremely sparse databases but was outperformed in other scenarios.

VII. CONCLUSION

This paper proposed TKU-PSO, a heuristic model based on particle swarm optimization for discovering top- k high-utility itemsets. TKU-PSO introduces several efficient strategies that are fundamental to the model's performance. First, we effectively reduced the number

of particle evaluations through fitness estimation and by utilizing explored candidates. Second, we introduced the concept of minimum solution fitness, which is employed in several stages of the algorithm to prune unpromising candidates. Finally, we revised the traditional population initialization and thus improved the model's ability to find optimal solutions in large search spaces. The experimental results show that our approach is superior in all tested datasets regarding execution time and accuracy. Most notably, THUI and TKO+ could not complete certain tests due to excessive runtimes, while TKU-PSO used less than 10 seconds to discover the correct solutions. In the other experiments, TKU-PSO was up to 63, 282, and 390 times faster than THUI, TKU-CE+, and TKO+, respectively. In addition, our model achieved an overall accuracy of 99.8% compared to 16.5% with TKU-CE+, and memory usage was the smallest on 4 of 6 datasets.

Although the proposed algorithm displays promising results, there are several opportunities for improving mining performance. Currently, the limiting factor to the speed of heuristics is the time required for candidate evaluations. Future work should thus focus on strategies that can reduce this cost, e.g., by omitting certain evaluations as proposed in this paper or designing a database projection that can be more efficiently scanned. Another possibility is to introduce parallel execution in the iterative stage of the algorithm such that it can perform concurrent evaluations. Regarding accuracy, we have demonstrated that population initialization can significantly impact the algorithm's ability to discover the correct solutions. Investigating whether this process can be further enhanced is thus an important topic to explore. Finally, the developed framework can also be adopted by other evolutionary techniques and extended for different utility mining problems.

REFERENCES

- [1] R. Agrawal, T. Imieliński, A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Record*, vol. 22, pp. 207–216, 1993, doi: <https://doi.org/10.1145/170036.170072>.
- [2] H. Yao, H. J. Hamilton, "Mining itemset utilities from transaction databases," *Data & Knowledge Engineering*, vol. 59, pp. 603–626, 2006, doi: <https://doi.org/10.1016/j.datak.2005.10.004>.
- [3] P. Fournier-Viger, J. Chun-Wei Lin, T. Truong-Chi, R. Nkambou, *A Survey of High Utility Itemset Mining*, pp. 1–45. Cham: Springer International Publishing, 2019.
- [4] H.-J. Choi, C. H. Park, "Emerging topic detection in twitter stream based on high utility pattern mining," *Expert Systems with Applications*, vol. 115, pp. 27–36, 2019, doi: <https://doi.org/10.1016/j.eswa.2018.07.051>.
- [5] H. Q. Vu, G. Li, R. Law, "Discovering highly profitable travel patterns by high-utility pattern mining," *Tourism Management*, vol. 77, p. 104008, 2020, doi: <https://doi.org/10.1016/j.tourman.2019.104008>.
- [6] M. Iqbal, M. N. Setiawan, M. I. Irawan, K. M. N. K. Khalif, N. Muhammad, M. K. B. M. Aziz, "Cardiovascular disease detection from high utility rare rule mining," *Artificial Intelligence in Medicine*, vol. 131, p. 102347, 2022, doi: <https://doi.org/10.1016/j.artmed.2022.102347>.
- [7] C. W. Wu, B.-E. Shie, V. S. Tseng, P. S. Yu, "Mining top-k high utility itemsets," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, p. 78–86.
- [8] D. Fogel, "What is evolutionary computation?," *IEEE Spectrum*, vol. 37, no. 2, pp. 26–32, 2000, doi: <https://doi.org/10.1109/6.819926>.
- [9] J. Kennedy, R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, vol. 4, 1995, pp. 1942–1948.
- [10] J. C.-W. Lin, L. Yang, P. Fournier-Viger, J. M.-T. Wu, T.-P. Hong, L. S.-L. Wang, J. Zhan, "Mining high-utility itemsets based on particle swarm optimization," *Engineering Applications of Artificial Intelligence*, vol. 55, pp. 320–330, 2016, doi: <https://doi.org/10.1016/j.engappai.2016.07.006>.
- [11] H. Rezk, J. Arfaoui, M. R. Goma, "Optimal parameter estimation of solar pv panel based on hybrid particle swarm and grey wolf optimization algorithms," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 145–155, 2021, doi: <https://doi.org/10.9781/ijimai.2020.12.001>.
- [12] K. K. Verma, B. M. Singh, "Deep multi- model fusion for human activity recognition using evolutionary algorithms," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 2, pp. 44–58, 2021, doi: <https://doi.org/10.9781/ijimai.2021.08.008>.
- [13] Y. Liu, W.-k. Liao, A. Choudhary, "A two-phase algorithm for fast discovery of high utility itemsets," in *Advances in Knowledge Discovery and Data Mining*, 2005, pp. 689–695.
- [14] C. F. Ahmed, S. K. Tanbeer, B.-S. Jeong, Y.-K. Lee, "Efficient tree structures for high utility pattern mining in incremental databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 12, pp. 1708–1721, 2009, doi: <https://doi.org/10.1109/TKDE.2009.46>.
- [15] V. S. Tseng, C.-W. Wu, B.-E. Shie, P. S. Yu, "Up-growth: An efficient algorithm for high utility itemset mining," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, p. 253–262.
- [16] U. Yun, H. Ryang, K. H. Ryu, "High utility itemset mining with techniques for reducing overestimated utilities and pruning candidates," *Expert Systems with Applications*, vol. 41, no. 8, pp. 3861–3878, 2014, doi: <https://doi.org/10.1016/j.eswa.2013.11.038>.
- [17] M. Liu, J. Qu, "Mining high utility itemsets without candidate generation," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 2012, p. 55–64.
- [18] P. Fournier-Viger, C.-W. Wu, S. Zida, V. S. Tseng, "Fhm: Faster high-utility itemset mining using estimated utility co-occurrence pruning," in *Foundations of Intelligent Systems*, 2014, pp. 83–92.
- [19] S. Krishnamoorthy, "Pruning strategies for mining high utility itemsets," *Expert Systems with Applications*, vol. 42, no. 5, pp. 2371–2381, 2015, doi: <https://doi.org/10.1016/j.eswa.2014.11.001>.
- [20] P. Wu, X. Niu, P. Fournier-Viger, C. Huang, B. Wang, "Ubp-miner: An efficient bit based high utility itemset mining algorithm," *Knowledge-Based Systems*, vol. 248, p. 108865, 2022, doi: <https://doi.org/10.1016/j.knsys.2022.108865>.
- [21] J. Liu, K. Wang, B. C. Fung, "Mining high utility patterns in one phase without generating candidates," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 5, pp. 1245–1257, 2016, doi: <https://doi.org/10.1109/TKDE.2015.2510012>.
- [22] S. Zida, P. Fournier Viger, J. C.-W. Lin, C.-W. Wu, V. Tseng, "EFIM: a fast and memory efficient algorithm for high-utility itemset mining," *Knowledge and Information Systems*, vol. 51, pp. 595–625, 2017, doi: <https://doi.org/10.1007/s10115-016-0986-0>.
- [23] H. Ryang, U. Yun, "Top-k high utility pattern mining with effective threshold raising strategies," *Knowledge-Based Systems*, vol. 76, pp. 109–126, 2015, doi: <https://doi.org/10.1016/j.knsys.2014.12.010>.
- [24] V. S. Tseng, C.-W. Wu, P. Fournier-Viger, P. S. Yu, "Efficient algorithms for mining top-k high utility itemsets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 54–67, 2016, doi: <https://doi.org/10.1109/TKDE.2015.2458860>.
- [25] Q.-H. Duong, B. Liao, P. Fournier-Viger, T.-L. Dam, "An efficient algorithm for mining the top-k high utility itemsets, using novel threshold raising and pruning strategies," *Knowledge-Based Systems*, vol. 104, pp. 106–122, 2016, doi: <https://doi.org/10.1016/j.knsys.2016.04.016>.
- [26] K. Singh, S. S. Singh, A. Kumar, B. Biswas, "TKEH: An efficient algorithm for mining top-k high utility itemsets," *Applied Intelligence*, vol. 49, no. 3, pp. 1078–1097, 2019, doi: <https://doi.org/10.1007/s10489-018-1316-x>.
- [27] J. Liu, X. Zhang, B. C. Fung, J. Li, F. Iqbal, "Opportunistic mining of top-n high utility patterns," *Information Sciences*, vol. 441, pp. 171–186, 2018, doi: <https://doi.org/10.1016/j.ins.2018.02.035>.
- [28] S. Krishnamoorthy, "Mining top-k high utility itemsets with effective threshold raising strategies," *Expert Systems with Applications*, vol. 117, pp. 148–165, 2019, doi: <https://doi.org/10.1016/j.eswa.2018.09.051>.
- [29] X. Han, X. Liu, J. Li, H. Gao, "Efficient top-k high utility itemset mining on massive data," *Information Sciences*, vol. 557, pp. 382–406, 2021, doi: <https://doi.org/10.1016/j.ins.2020.08.028>.
- [30] M. Ashraf, T. Abdelkader, S. Rady, T. F. Gharib, "TKN: an efficient approach for discovering top-k high utility itemsets with positive or negative profits," *Information Sciences*, vol. 587, pp. 654–678, 2022, doi: <https://doi.org/10.1016/j.ins.2021.12.024>.
- [31] C. Zhang, Z. Du, W. Gan, P. S. Yu, "Tkus: Mining top-k high utility

sequential patterns,” *Information Sciences*, vol. 570, pp. 342–359, 2021, doi: <https://doi.org/10.1016/j.ins.2021.04.035>.

- [32] W. Song, C. Zheng, C. Huang, L. Liu, “Heuristically mining the top-k high-utility itemsets with cross-entropy optimization,” *Applied Intelligence*, pp. 1–16, 2021, doi: <https://doi.org/10.1007/s10489-021-02576-z>.
- [33] J. C.-W. Lin, L. Yang, P. Fournier-Viger, T.-P. Hong, M. Voznak, “A binary pso approach to mine high-utility itemsets,” *Soft Computing*, vol. 21, no. 17, p. 5103–5121, 2017, doi: <https://doi.org/10.1007/s00500-016-2106-1>.
- [34] W. Song, C. Huang, “Mining high utility itemsets using bio-inspired algorithms: A diverse optimal value framework,” *IEEE Access*, vol. 6, pp. 19568–19582, 2018, doi: <https://doi.org/10.1109/ACCESS.2018.2819162>.
- [35] W. Fang, Q. Zhang, H. Lu, J. C.-W. Lin, “High-utility itemsets mining based on binary particle swarm optimization with multiple adjustment strategies,” *Applied Soft Computing*, vol. 124, p. 109073, 2022, doi: <https://doi.org/10.1016/j.asoc.2022.109073>.
- [36] S. Kannimuthu, K. Premalatha, “Discovery of high utility itemsets using genetic algorithm with ranked mutation,” *Applied Artificial Intelligence*, vol. 28, no. 4, pp. 337–359, 2014, doi: <https://doi.org/10.1080/08839514.2014.891839>.
- [37] Q. Zhang, W. Fang, J. Sun, Q. Wang, “Improved genetic algorithm for high-utility itemset mining,” *IEEE Access*, vol. 7, pp. 176799–176813, 2019, doi: <https://doi.org/10.1109/ACCESS.2019.2958150>.
- [38] J. M.-T. Wu, J. Zhan, J. C.-W. Lin, “An aco-based approach to mine high-utility itemsets,” *Knowledge-Based Systems*, vol. 116, pp. 102–113, 2017, doi: <https://doi.org/10.1016/j.knsys.2016.10.027>.
- [39] W. Song, C. Huang, “Discovering high utility itemsets based on the artificial bee colony algorithm,” in *Advances in Knowledge Discovery and Data Mining - 22nd Pacific-Asia Conference*, vol. 10939, 2018, pp. 3–14.
- [40] W. Song, J. Li, C. Huang, “Artificial fish swarm algorithm for mining high utility itemsets,” in *Advances in Swarm Intelligence - 12th International Conference*, vol. 12690, 2021, pp. 407–419.
- [41] M. S. Nawaz, P. Fournier-Viger, U. Yun, Y. Wu, W. Song, “Mining high utility itemsets with hill climbing and simulated annealing,” *ACM Transactions on Management Information Systems*, vol. 13, no. 1, 2021, doi: <https://doi.org/10.1145/3462636>.
- [42] R. Rubinstein, “The cross-entropy method for combinatorial and continuous optimization,” *Methodology and computing in applied probability*, vol. 1, no. 2, pp. 127–190, 1999, doi: <https://doi.org/10.1023/A:1010091220143>.
- [43] P. Fournier-Viger, J. C.-W. Lin, A. Gomariz, T. Gueniche, A. Soltani, Z. Deng, H. T. Lam, “The SPMF open-source data mining library version 2,” in *Machine Learning and Knowledge Discovery in Databases - European Conference*, vol. 9853, 2016, pp. 36–40.

multiple types of data mining algorithms. He also serves as the Editor-in-Chief of the International Journal of Data Science and Pattern Recognition. He is the IET Fellow, senior member for both IEEE and ACM.



Simen Carstensen

Simen received his B.Sc. in Computer Science (2020) from the University of Bergen (UiB), and a subsequent M.Sc in Software Engineering (2022) from Western Norway University of Applied Sciences (HVL). With a passion for optimizing computer systems through algorithms, his expertise is primarily centered around data mining, meta-heuristics, and machine learning. Presently, Simen applies

his skills as a developer at Dataloy Systems, focusing on building efficient and streamlined solutions for the shipping industry.



Jerry Chun-Wei Lin

Jerry Chun-Wei Lin received his Ph.D. from the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan in 2010. He is currently a full Professor with the Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway. He has published more

than 300 research articles in refereed journals (IEEE TKDE, IEEE TCYB, ACM TKDD, ACM TDS, ACM TMIS) and international conferences (IEEE ICDE, IEEE ICDM, PKDD, PAKDD). His research interests include data mining, soft computing, artificial intelligence and machine learning, and privacy preserving and security technologies. He is also the project co-leader of well-known SPMF: An Open-Source Data Mining Library, which is a toolkit offering

Prediction of COVID-19 Using a Clinical Dataset With Machine Learning Approaches

A. Suruliandi¹, R. Ame Rayan^{1*}, S. P. Raja²

¹ Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, (India)

² School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, (India)

* Corresponding author: amerayanld@gmail.com

Received 20 August 2023 | Accepted 11 July 2024 | Early Access 29 January 2025



ABSTRACT

COVID-19 is an infectious disease that spreads quickly from person to another. The pandemic, which spread worldwide over time, presents huge risks in terms of blood clotting, breathing problems and heart attacks, sometimes with fatal consequences if not detected early. The PCR test, CT scans, X-rays, and blood tests are methods commonly employed to detect the disease, though the PCR test is, without question, considered the gold standard. The American Center for Disease Control and Prevention (CDC) reports that the PCR has an 80% accuracy rate. An alternative to the PCR is clinical data, which is less expensive, easy to collect, and offers better accuracy. Machine learning, with its rich feature selection and classification methods, helps detect COVID-19 at the earliest stages, using clinical test results. This research proposes a clinical dataset and offers a comparative analysis of feature selection and classification algorithms for detecting COVID-19. Filter-based feature selection methods such as the ANOVA-F, chi-square, mutual information and Pearson correlation, along with wrapper-based methods such as Recursive Feature Elimination (RFE) and Sequential Forward Selection (SFS) were used to choose a subset of features from the feature set. The selected features were thereafter applied to the Support Vector Machine (SVM), Naïve Bayes, K-NN (K-Nearest Neighbor) and Logistic Regression(LR) classification algorithms to detect Coronavirus Disease. The experimental results of the comparative study show that the clinical dataset provides better accuracy at 94.8%, with mutual information and the SVM classifier.

KEYWORDS

Blood Samples, Classification, COVID-19 Prediction, Feature Selection, Machine Learning.

DOI: 10.9781/ijimai.2025.01.003

I. INTRODUCTION

In December 2019, the novel coronavirus caused a public health crisis which spread rapidly worldwide. The disease is transmissible, striking healthy individuals who come in contact with droplets from an infected individual [1]. The infected person is sometimes asymptomatic, while others develop symptoms like cough, fever, shortness of breath and body pain, as well as loss of taste and smell. The Reverse Transcription Polymerase Chain Reaction (RT-PCR) has been followed as the gold standard in diagnosing COVID-19 [2]. Notwithstanding its popularity, the test has certain intrinsic flaws in that it consumes more time, expensive, requires specially designed laboratory devices, and has a false negative rate of 20% [3]. Blood tests, X-rays, CT scans and breath sound analysis have been used as alternative procedures in COVID-19 diagnosis. Even though positive results are obtained using chest X-ray and CT scans images based on machine learning [4], the downside of these tests is exposure to high doses of radiation. Given that recent studies have shown that the blood features of COVID-19 patients change dramatically [5]–[10], hence early detection of the virus can

be done by recognizing and working with these parameters. The blood tests results are ready in quick time and are relatively cheaper than other tests.

A decision support system is most useful in predicting COVID-19 using clinical data in the early stages so appropriate decisions can be made in good time. Clinical data includes biochemical parameters, obtained through blood tests that are made easily available in little time. The parameters include C-reactive protein (CRP), lymphocytes, DC:Neutrophils and D-Dimer, among others, which show changes due to coronavirus infection. As a result, the most common clinical findings, such as biochemical and hematological parameters, play an important role in COVID-19 preliminary screening [11].

Researchers in Artificial Intelligence use machine learning as a tool to assist healthcare workers diagnose disease. Machine learning classification and clustering algorithms give the best results when it comes to building such decision support systems. Machine learning provides algorithms that handle large datasets in a minimum runtime by selecting appropriate attributes. Further, it provides excellent

Please cite this article as:

A. Suruliandi, R. Ame Rayan, S. P. Raja. Prediction of COVID-19 Using a Clinical Dataset With Machine Learning Approaches, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 4, pp. 82-98, 2025, <http://dx.doi.org/10.9781/ijimai.2025.01.003>

detection methods [12]. Machine learning models speed up information analysis and help make efficient disease prediction decisions. A model designed using machine learning recognizes patterns in blood samples and uses them to diagnose COVID-19. The objectives of this paper are (i) to propose a new clinical dataset to predict whether the person is affected by COVID-19 or not, and (ii) to find an optimal feature selection method and classifier for COVID-19 prediction.

II. RELATED WORK

An analysis of the literature highlights similar work on COVID-19 prediction using blood test datasets. A tool was designed by Wu et al. [13] using the random forest algorithm to predict COVID-19. A total of 253 samples of data were collected for this purpose from 169 suspected patients. Each instance of data had 49 parameters, with 24 and 25 of those relating to the hematological and biochemical, respectively. In all, 11 parameters were extracted with the help of random forest algorithm. The overall performance of the tool in terms of accuracy in COVID-19 prediction was measured at 95.95%. Bastug et al. [14] undertook a comprehensive analysis of laboratory and clinical attribute for detecting COVID-19. The severity of the illness is predicted by training the model with the information from 191 coronavirus affected patients, admitted at an Ankara city hospital. In all, 29 blood routine parameter features were statistically analyzed. Kolmogorov-Smirnov test was used to check the normality of variables and to predict disease severity binary logistic regression was applied.

Brinati et al. [15] developed two machine learning models for COVID-19 detection. The study collected 279 blood samples from 177 COVID-19 positive and 102 COVID-19 negative individuals. The missing values were handled using Multivariate Imputation by Chained Equation (MICE) and feature Importance was used to select the best features. Five different machine learning algorithms, including the extremely randomized trees, logistic regression, decision tree, k-nearest neighbors, Naïve Bayes and random forest were compared to detect COVID-19. Of these, the two models designed using random forest algorithm outperformed with the accuracy of 82% and 86% respectively. Kukar et al. [16] used data from the University Medical Center in Ljubljana, Slovenia, to train a machine learning model to detect COVID-19. Blood samples from 5333 patients with viral and bacterial infections and 160 COVID-19 positive patients were included in the dataset. Feature selection was carried out using the feature importance scoring feature of the XGBoost machine learning algorithm, and the predictive model designed using the algorithm yielded an AUC of 0.97 in detecting COVID-19. Chadaga et al. [17] used data from Brazil's Albert Einstein Hospital to build a model for COVID-19 diagnosis. This model used SMOTE balancing technique to balance the dataset through oversampling, along with correlation analysis and feature importance to select the best features. The random forest, k-nearest neighbors, logistic regression, and XGBoost classifiers were compared using this dataset, and the best accuracy (92%) was produced by random forest algorithm.

Aljame et al. [18] proposed the "ER-CoV" machine learning model to predict the incidence of COVID-19 using hematological and demographic parameters. Data collected from 5644 patients of the Albert Einstein Hospital, Brazil, were preprocessed using the KNNImputer algorithm to handle null values and the SMOTE to balance the dataset. The SHAP technique was used to select 18 features from a total of 108. The proposed model has two level classifiers. The first level had the random forest, logistic regression and extra trees classifiers, and the output from this level was given as input to the second-level extreme gradient boosting classifier to detect COVID-19. The proposed model achieved 99.88% overall accuracy. The authors of

[19] proposed a model using five ML algorithms such as gradient boost trees, SVM, logistic regression, neural networks, and random forest for the diagnosis of COVID-19. A dataset was created using the data collected from Brazil's Albert Einstein Hospital. The dataset contains 235 blood samples with 102 confirmed cases of COVID-19. From the dataset, 15 relevant characteristics were chosen for research. In this study, the SVM produced the best classification results with very little significance, with AUC, sensitivity, and specificity of 85%, 68%, and 85%, respectively, when compared to previous work. A study [20] analyzed and applied six state-of-the-art methods like the SVM, MLP, NB, RT, Bayesian Networks (BN), and RF to a dataset from the Brazil's Albert Einstein Hospital which consists of 564 samples, including 559 COVID-19 positive samples. The SMOTE was used for oversampling due to limited size of the dataset. Two PSO-based algorithms, the evolutionary search algorithm, and a manual method were all used for feature selection. The BN model performed the best overall, achieving accuracy, precision, specificity, and sensitivity values of 95.159%, 93.8%, 93.6%, and 96.8%, respectively.

Almansoor and Hewahi [21] collected Kaggle data with patient information from the Brazil's Albert Einstein Hospital, containing 5644 instances and 111 features. Data preprocessing was carried out using a one-sided selection technique to balance the data. The SVM, AdaBoost, random forest and k-nearest neighbour classifiers were used to detect COVID-19. Cabitza et al. [22] compared the performance of their model using the random forest, logistic regression, k-NN, SVM and Naïve Bayes algorithms. Three types of datasets, namely, the CBC, OSR, and a COVID-19-specific dataset were utilized. It was observed that the random forest and SVM performed the best with 88% accuracy for the OSR dataset, while the k-NN and SVM outperformed other algorithms on the COVID-19 specific dataset. The CBC dataset produced good results with the k-NN algorithm.

Akhtar et al. [23] used various machine learning algorithms like the k-NN, SVM, Naïve Bayes, multi-layer perceptron and decision tree to detect COVID-19 using the CBC dataset uploaded on the Kaggle website. The CBC dataset contains the CBC parameters of 5644 patients. Performance-wise, the multi-layer perceptron outclassed other algorithms. Abayomi-Alli et al. [24] introduced an ensemble learning model for COVID-19 detection using blood test samples. They combined custom convolutional neural networks (CNN) with 15 supervised machine learning algorithms. This ensemble model, incorporating DNN and ExtraTrees, achieved a remarkable accuracy of 99.28% and an AUC of 99.4% on the San Raffaele Hospital dataset, outperforming other COVID-19 diagnostic methods. Gong et al. [25] present a methodology for achieving explainable AI-driven rapid COVID-19 diagnosis. They employed ensemble learning algorithms to analyze data collected from 1,737 participants hospitalized at San Raphael Hospital during the period of February to May 2020. The study applied four distinct ensemble learning algorithms, namely random forest, adaptive boosting, gradient boosting decision tree (GBDT), and extreme gradient boosting (XGBoost). Notably, the GBDT model demonstrated superior performance, achieving an accuracy of 86.4% in effectively distinguishing COVID-19 patients from the control group. Roy and Singh [26] introduced a framework employing the weighted average of predictive accuracy from individual transfer learning models, including ResNet50V2, DenseNet201, and InceptionNetV3. The framework demonstrated exceptional performance in detecting COVID-19 from Chest X-ray images, achieving an impressive F1-score of 0.997.

Andueza et al. [27] used ARIMA and SARIMA Machine Learning models to predict the impact of COVID-19 on tobacco sales in Spain (January 2020 to December 2021) in euros, packs, and per capita packs. The study highlights a significant decline in cigarette sales, particularly in provinces popular among tourists and those sharing

borders with France. Sales during border closures were up to 66.74% lower than the initial forecasts, emphasizing the notable impact of COVID-19 restrictions on provincial tobacco sales in Spain. This suggests a disruption in the typical patterns of tourism and cross-border purchases between Spain and France, as well as Spain and Gibraltar. Cowley et al. [28] suggested a novel approach that integrates the outcomes of supervised random forest classification with unsupervised clustering to forecast patient risk. The model demonstrated superior performance, achieving an accuracy of 92%.

A. Motivation and Justification

COVID-19, a communicable disease that has spread throughout the globe, has such common symptoms that it has facilitated the publication of numerous open source clinical datasets. The literature review makes it plain that clinical parameters help in early screening of the disease. A person infected by the coronavirus shows variations in blood components. Given that the C-Reactive Protein (CRP), DC:Neutrophils, D-Dimer, and the lymphocytes vary rapidly from their normal values, they help identify infected individuals with early screening and have thus motivated the creation of an open source clinical dataset. The proposed clinical dataset contains COVID-19 patients' blood test results, and it is anticipated that it will help researchers develop a tool for an initial screening of the disease.

It is clear from the literature survey that the datasets used for the research have missing values and are imbalanced. The two issues are addressed by machine learning techniques such as the Multivariate Imputation by Chained Equation (MICE), KNNImputer, and SMOTE. The two factors above have motivated the creation of an open source clinical dataset with balanced data and with no missing values. Secondly, it is seen that feature selection is vital to improved performance. The two types of feature selection algorithm that is widely used are filter and wrapper based methods. The filter based method selects features with a score greater than the threshold. The wrapper method selects a subset of features for classification, following which the subset with the best accuracy is selected as the best feature set. Reducing the number of features by selecting optimal ones helps improve the performance of the model. This research is justified in that it carries out numerous experiments, using the proposed clinical dataset, to perform a comparison in the accuracy of the classifier with and without feature selection. Machine learning algorithms such as the ANOVA-F, chi-square, mutual information, Pearson coefficient, SFS and RFE are used for feature selection.

The literature review revealed that machine learning classifiers such as the random forest, XGBoost, logistic regression, extra trees, SVM, Naïve Bayes, and multilayer perceptron help predict the disease most accurately. This research uses the Naïve Bayes, SVM, k-NN and logistic regression to predict the disease using the features selected from the feature selection algorithms mentioned above. To summarize, machine learning algorithms may be used in prediction by training the dataset and incorporating the given input data with the trained data for classification. Most healthcare applications, therefore, use machine learning approaches for prediction.

Classification techniques such as the logistic regression, SVM, k-NN and Naïve Bayes are used to classify the selected features. The findings indicate that the performance of the classifier is enhanced by only using selected features that are picked following the application of feature selection. The prediction model, built with the selected features and classification algorithms in machine learning, produces good accuracy. This research work undertakes a comparative analysis of several feature selection and classification algorithms to discover the most effective feature set and classifier respectively, for detecting COVID-19 using the proposed clinical dataset.

B. Outline of the Work

The overall working of the research process is shown in Fig. 1. Firstly, the dataset is preprocessed to handle missing values, eliminate redundant values and convert categorical values into numerical values. Then the dataset is checked for outliers, the identified outliers are removed, and the dataset is balanced using SMOTE algorithm. Secondly, from the preprocessed dataset, significant features are selected using feature selection algorithm. Thirdly, the data in the selected features are subject to several classification techniques to identify the persons affected by COVID-19. Finally, based on the performance metrics of various classification techniques, best feature selection and classification algorithm is selected.

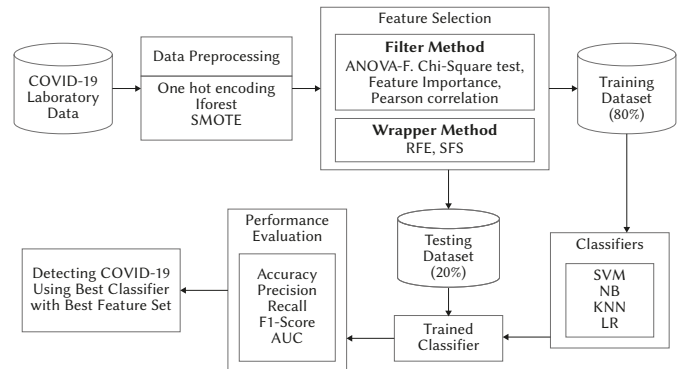


Fig. 1. Outline of the Work.

C. Organization of the Paper

The rest of this research paper is organized as follows. Section 3 discusses the methodology of the model. Section 4 describes the procedure for COVID-19 prediction. Section 5 presents the findings of the experiments and discusses them, while Section 6 concludes the paper and offers directions for future research.

III. METHODOLOGY

A. Proposed Dataset Construction

There are, currently, only a few clinically available COVID-19 datasets which, for the most part, however, cannot be used by researchers directly, given that most of the features presented therein have missing values. Data need to be preprocessed. When missing data are handled during preprocessing, a particular feature is either dropped or filled by using statistical formulae. Such procedures, then, fail to produce accurate results. In most clinical datasets, information on positive and negative COVID-19 patients is not balanced. Therefore, a new clinical dataset was constructed using information on 2000 patients with COVID-19 symptoms, all of whom had taken a blood test between August 2020 and August 2021, at a private hospital in Thoothukudi, Tamil Nadu, India. Patient data privacy is maintained by excluding the patient's name and any other personal information. Instead, a fictitious patient number is linked to the collected blood sample data. Based on the values of the 27 features in the blood test, the final result was noted as COVID-19 positive or negative. Table I provides a description of the clinical dataset. Following the acceptance of the paper, the entire dataset will be made accessible publicly via the link: <https://github.com/merviname/COVID-19>.

B. Data Preprocessing

The data acquired are preprocessed by checking for missing values and duplicate entries. The dataset contains no missing value. The duplicate entries in the dataset are removed. The dataset contains

TABLE I. COVID-19 CLINICAL DATASET DESCRIPTION

S.No	Field	Data Type	Description	Normal Range values
1	AGE	Numerical	Patient Age	>0
2	GENDER	Categorical	Patient Gender	M-Male F-Female
3	HB	Numerical	It is the measurement of Hemoglobin in blood. COVID-19 patients have a decrease in HB which indicates the low oxygen carrier in blood.	Male-13.5– 17.0 g/dl Female-12.0–15.5 g/dl
4	TC	Numerical	It stores the count of white blood cells. Patients affected by COVID-19 have a significant increase of white blood cells.	4000-11000 cells
5	DC:NEUTROPHILS	Numerical	It stores the count of neutrophils in the blood. Patients affected by COVID-19 have increase in neutrophil count.	40-65 %
6	LYMPHOCYTES	Numerical	It stores the lymphocytes count. The lymphocytes count decrease in COVID-19 patients.	30-50 %
7	EOSINOPHILS	Numerical	It stores the Eosinophil count in the blood to measure the allergic disease, infections etc.COVID-19 patients has a decrease in eosinophil count.	100-400 cells/mL
8	MONOCYTES	Numerical	It is a kind of white blood cell which fight against disease and infections. COVID-19 patients has a decrease in Monocytes count.	200-800 /mL
9	BASOPHILS	Numerical	It is used to measure the allergic reaction. COVID-19 patients has a lower basophils count.	0-300 /mL
10	ESR (60 MIN)	Numerical	Erythrocyte Sedimentation Rate-It is used to measure and identify the inflammation. There is an increase of ESR in COVID-19 patients.	Male : 0-17 mm/hour Female : 1-25 mm/hour Children : 0-10 mm/hour
11	PC	Numerical	It stores the average platelets count in the blood. COVID-19 patients has a lower platelet count.	150 - 410 thousands/cmm
12	PCV	Numerical	Packed Cell Volume - It stores the Red Blood Cells proportion in blood. Patients affected by COVID-19 have a decrease in PCV.	Male : 40-52% Female : 35- 47%
13	MCV	Numerical	Mean Corpuscular Volume – It stores the size of Red Blood Cell. Patient affected by COVID-19 has low MCV value.	80-95 fL
14	MCH	Numerical	Mean Corpuscular Hemoglobin – It stores the Hemoglobin amount in Red Blood Cell. Patient affected by COVID-19 has low MCH value.	27.5 - 33.2 pg/cell
15	MCHC	Numerical	Mean Corpuscular Hemoglobin Concentration – It stores the average amount of hemoglobin in the group of Red Blood Cells. Patient affected by COVID-19 has low MCHC value.	32 – 36 g/dL
16	RBC	Numerical	It stores the number of Red Blood cells in the blood. In severe COVID-19 affected patients shows low RBC.	Male: 4.7 – 6.1 million cells/microliter Female: 4.2– 5.4 million cells/microliter
17		Numerical	Red Cell Distribution Width – It stores the variances in the size and volume of Red Blood Cells. Severe COVID-19 affected patient shows high RDW-CV count.	Male : 11.8 – 14.5 % Female : 12.2 – 16.1 %
18	RBS	Numerical	Random Blood Sugar Test. It stores the level of Blood Sugar of a non-fasting person. COVID-19 patients will have increase in the blood sugar level.	< 140 mg/dL
19	UREA	Numerical	It stores the amount of Urea in the Blood sample. Urea level in blood is high for COVID-19 patients.	6 – 24 mg/dL
20	CREATININE	Numerical	It stores the measure of creatinine in the blood sample. Creatinine in blood sample has an increase in COVID-19 patients.	Male :0.74 – 1.35 mg/dL Female : 0.59 – 1.04 mg/dL
21	CRP	Numerical	It stores the measure of C-reactive protein in the blood. There is significant increase in the CRP value for COVID-19 patients.	0-5 mg/L
23	D-DIMER	Numerical	It stores the measure of protein fragments of blood clots floating in the blood. COVID-19 patients have a higher D-Dimer value.	< 500 mg/mL
24	LDH	Numerical	It stores the amount of Lactate Dehydrogenase in the blood. There is significant increase in the LDH amount for COVID-19 patients.	125 – 343 U/L
25	DIRECT BILLIRUBIN	Numerical	It stores the measure of conjugated bilirubin. There is an increase in the direct Bilirubin value for COVID-19 patients.	0 – 0.3 mg/dL
26	BILLIRUBIN T	Numerical	It stores the sum of Direct and Indirect Bilirubin. There is an increase in the Indirect Bilirubin value for COVID-19 patients.	0.1–1.2 mg/dL
27	INDIRECT BILLIRUBIN	Numerical	It stores the measure of unconjugated bilirubin There is an increase in the Indirect Bilirubin value for COVID-19 patients.	0.2–0.8 mg/dL
28	SGOT	Numerical	Serum-Glutamic-Oxaloacetic-Transaminase – It stores the measure of enzyme found in liver, heart and other tissues. There is an increase in the SGOT value for COVID-19 patients.	8–45 units/litre

categorical and numeric values. Machine learning algorithms work well with numerical data. Hence, to convert categorical values into numerical values, one-hot encoding is used. The outliers in the dataset are removed using iForest algorithm and then the dataset is balanced using SMOTE algorithm.

1. One-Hot Encoding

Features with string values refer to categorical data, while most machine learning algorithms work with numerical values. Hence, these categorical values have to be mapped with numerical values. This conversion helps the algorithm for better prediction [29]. In this study, the categorical feature, 'gender' is converted into a numerical feature by creating two columns, Gender_1 and Gender_2 by using one-hot encoding method.

2. iForest

Anomalies in a dataset differ from normal records both in terms of quantity and quality. Removing these outliers can significantly enhance the performance of a classification model. In the context of this study, the Isolation Forest (iForest) [30] technique was employed to identify and eliminate outliers from the proposed COVID-19 clinical dataset. iForest identifies outliers by calculating the average path lengths for instances within its tree structures, with outliers being instances having notably shorter average path lengths.

iForest demonstrates efficient performance when used with a relatively small subsample size and an appropriate number of trees. The 'contamination' parameter serves the purpose of specifying the proportion of outliers present in the dataset. For this particular study, the chosen parameter configuration led to the detection of 24 outliers in the dataset mentioned above. After the removal of these outliers, the subsequent step involved addressing the issue of dataset imbalance. Imbalanced data can significantly impact the performance of a classification model, especially during training. Imbalanced data often causes the classification model to exhibit bias toward the majority class, leading to an increased occurrence of both false positives and false negatives. This, in turn, diminishes the overall performance of the classification model. Therefore, to enhance the performance, the proposed classification model balanced the COVID-19 data.

The clinical dataset proposed for this study displayed a significant imbalance, consisting of 997 COVID-positive cases and 999 being COVID-negative cases. This stark imbalance tilted the dataset heavily toward negative cases. To address this imbalance, the proposed model employed the Synthetic Minority Over-sampling Technique (SMOTE) to randomly generate minority class instances, effectively oversampling the minority class and rebalancing the dataset. Then, the entire dataset was randomly split into an 80% training set and a 20% test set.

C. Feature Selection

The columns/attributes in the dataset are termed features and only essential ones are needed to train an optimal model. Feature selection, which is the process of choosing essential features, is critical to building a machine learning model because it reduces data redundancy and thus maximizes the model's performance. The objectives of feature selection techniques are (i) to reduce the model's complexity by removing irrelevant features, (ii) to help the machine learning algorithm train a model faster, and (iii) to avoid overfitting by reducing the dimensions [31]. Based on its interaction with the classifier, the feature selection algorithm is divided into three types they are: (i) filter method, (ii) wrapper method and (iii) embedded method. This study makes use of filter and wrapper methods.

1. Filter Methods

Statistical techniques are used in the filter method to assess the dependence between the input variable and the target variable.

Statistical measures such as Fisher score, mutual information, chi-square test, correlation coefficient and variance threshold identify important features [32]. The techniques calculate the scores based on variance, correlation, consistency and distance, depending on the data's intrinsic properties. Thereafter, the features are ranked from best to worst, based on the said scores [33]. Fig. 2. shows the operation of the filter method. This paper employs the following four filter-based feature selection methods: (i) ANOVA-F (ii) chi-square (iii) mutual information (iv) Pearson correlation.

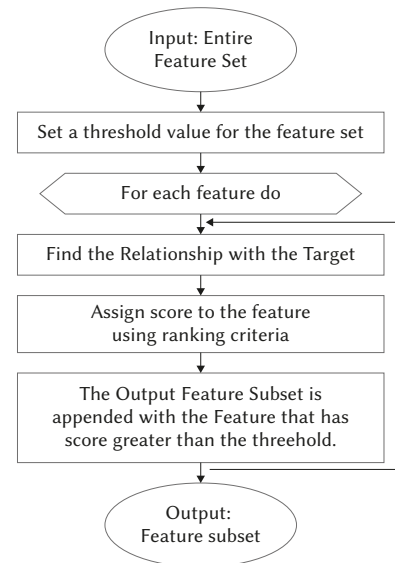


Fig. 2. Filter Method [41].

a) ANOVA-F:

ANOVA (Analysis of Variance) analyses each feature individually to examine the feature - target relationship. Features with a tenuous relationship with the target are eliminated. The F-Test is a statistical function that computes the ratio of the variance values. The variance denotes the dispersal measure of the data points from the mean. The ANOVA-F is used for the numerical input variable with the classification target variable [34]. Based on the test results, the best features with a high F-statistic score are selected. Features with a low F-statistic score, which are independent of the target variable, are removed from the dataset.

b) Chi-Square:

The chi-square test identifies attributes that are highly dependent on the target variable. It measures dependencies by examining the deviation of the expected count from the observed count. The chi-square value is small when the observed count is close to the expected count, indicating that the input feature is independent of the response. The higher chi-square value shows that the dependencies between the feature and response is high [35]. The chi-square feature selection algorithm selects features by calculating the chi-score and the p-value. The most significant features have a high chi-square score and a low p-value.

c) Mutual Information:

Mutual Information calculates the entropy for each feature with reference to the target feature [36]. Mutual information is calculated for each independent feature, following which the features are ranked, based on the calculated information gain for each feature. A threshold is set for selecting features with information gain above the threshold value. Mutual information thus helps find the most useful features that differentiate the target class.

d) Pearson Correlation:

Pearson correlation constructs a correlation matrix that measures the linear association between two features. The values in the matrix range from -1 to 1. Values closer to -1 and 1 indicate strong negative correlation and positive correlation, respectively. Values closer to 0 indicate weak correlation, while features with a value of 0 have no correlation [37]. A threshold is set to select the best features, and those with a higher score than the threshold is selected while others are removed from the dataset.

2. Wrapper Methods

The wrapper-based feature selection technique selects the best feature subset by producing a number of candidate feature subsets whose accuracy is evaluated using a classification algorithm. The best feature set is defined as the feature subset with the highest accuracy [38]. Fig. 3 shows the working of the wrapper-based feature selection method. The wrapper method such as Recursive Feature Selection (RFE) and Sequential Forward Selection (SFS) are used in this research.

a) Recursive Feature Elimination (RFE):

The Recursive Feature Elimination (RFE) algorithm first determines the most significant features and subsequently removes the least important ones, one at a time, in each iteration. The features are eliminated repeatedly until an optimal threshold is obtained from the classification algorithm. The final feature set obtained is the best [39]. Each feature is ranked using the `rfe_ranking` and features with '1' in the `rfe_ranking` column are selected for classification.

b) Sequential Forward Selection (SFS):

The Sequential Forward Selection (SFS) algorithm initially has an empty set of features, with features added on to the feature set at each iteration. The best feature set is obtained when the iteration yields a reduced misclassification rate [40]. The average score for each feature subset is calculated. Initially, the average score starts with a single feature and, at each iteration, another feature is added to the subset. The feature subset with the highest average score is selected as the best, and the features within it are selected for classification.

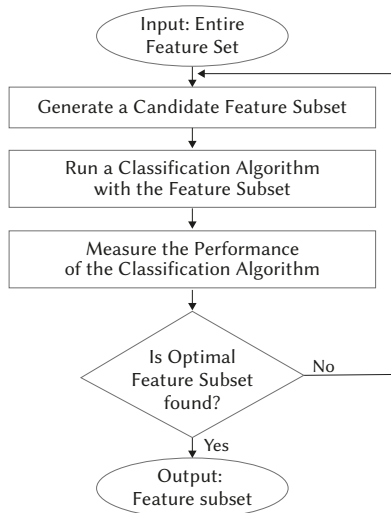


Fig. 3. Wrapper Method [41].

D. Classifiers

A classifier is a machine learning algorithm that can be used to identify the class of given input data. It takes the input data and outputs discrete class labels that define a set of possible classes. Classifiers, once trained with machine learning classification algorithms, can be used to

make predictions on new data points and identify the class to which a training set belongs [42]. Several machine learning classification algorithms are used in this paper to predict COVID positive cases based on patients' blood test results.

Supervised machine learning models such as the SVM, Naive Bayes, KNN and logistic regression are applied for learning the preprocessed data after selecting the best features. The dataset is divided into training and testing data in 80:20 ratio. The classifier algorithm is used to train the model using training set. After that 20 percent of the set is used as testing data. The implementation process of the models used in this study follow below.

1. Support Vector Machines

The support vector machine is a statistics-based supervised machine learning algorithm that is used for classification and regression [43], which enables it to predict COVID-19 with its features. The SVM creates a decision boundary known as the hyperplane which differentiates between COVID-19 positive and negative classes. The selected features from the clinical dataset are trained using the SVM algorithm. The training process results in a set of support vectors and a decision boundary. A predictive analysis is carried out using $w^*x_i - c = +1$ and $w^*x_i - c = -1$ (where 'w' is the vector which is normal to the hyperplane and 'c' the offset) by dividing the points on the hyperplane [44]. The SVM classifies the given new input vectors by calculating the distance from the decision boundary. The distance (d) from the point (a₀, b₀) to the line $Mx + Ny + O$ is calculated using Eq. 1.

$$d = \frac{|Ma_0 + Nb_0 + O|}{\sqrt{M^2 + N^2}} \quad (1)$$

Similarly, the distance between the hyperplane $w^T(\Phi(x)) + c$ and the given vector $\Phi(a_0)$ is given by Eq 2.

$$d_H(\Phi(a_0)) = \frac{|w^T(\Phi(a_0)) + b|}{\|w\|_2} \quad (2)$$

where 'w' is the vector that is normal to the hyperplane, 'b' the offset and $\|w\|_2$ the length of w in the Euclidean norm. $\|w\|_2$ is given by $\|w\|_2 = \sqrt{w_1^2 + w_2^2 + w_3^2 + \dots + w_n^2}$.

For better accuracy using the SVM, the algorithm maximizes the distance and gives space to the hyperplane. Hence, to maximize the minimum distance, Eq. 3 is used.

$$w^* = \operatorname{argmax}[\min d_H(\Phi(a_n))] \quad (3)$$

If a point is substituted in the hyperplane equation ($w^*x + c > 0$) and is greater than zero, then the given data is COVID-19 positive. If a point is substituted in the hyperplane equation ($w^*x + c < 0$) and is less than zero then the given data is COVID-19 negative.

Pseudocode for the Support Vector Machine

Input: D = [X, Y]; X (dataset with m features), Y (class labels)

Output: Test case class

1. Initialize the model with random values for the weights, w.
2. Split the dataset, D, into two parts: a training set, T, and a testing set, T1.
3. For each training data, xi, from dataset T:
 - 3.1 Compute the margin, $y_i(w) = w^T x_i$.
 - 3.2 If the margin is greater than 1, add xi to the support vector set, S.
4. Find the optimal weights, w*, by solving the quadratic optimization problem, subject to $y_i(w) \geq 1$ for all i in S.
5. Return the support vector set, S, and the optimal weights, w*.
6. Assign the test data as positive if the $y_i(w) \geq 1$, otherwise classify it as negative.

2. Naïve Bayes:

The Naïve Bayes algorithm is a statistical supervised machine learning algorithm that predicts class membership using probability. The algorithm works well for small datasets but even better for large ones, offering high accuracy and speed [45].

Pseudocode for the Naïve Bayes [46]

Input: D = [X, Y]; X (dataset with n features), Y (class labels)

F = (f1, f2, f3, ..., fn) // features in the testing dataset

Output: Test case class

1. Split the dataset, D, into two parts: a training set, T, and a testing set, T1.
2. Read the training dataset, T.
3. Repeat to Compute the probability of fi using the Gauss density equation in each class until the probability of all predictor variables (f1, f2, f3, ..., fn) has been calculated
4. Compute the likelihood for each class.
5. Select the greatest likelihood.

The Naïve Bayes algorithm is based on Bayes' theorem, written as in Eq. 4

$$P(C|D) = \frac{P(D|C) * P(C)}{P(D)} \quad (4)$$

where P(C) is the prior probability denoting the probability of occurrence C and P(D) the marginal probability denoting the probability of occurrence D. The probability values are independent and do not refer to each other. P(C|D) is known as the posterior probability which represents the probability of occurrence of C, given that D has occurred. This algorithm does not depend on other parameters and uses Eq. 5 to predict COVID-19.

$$P(C|D1 \dots Dn) \propto \prod_{i=1}^n P(D_i|C) P(C) \quad (5)$$

Eq. 6 below is used to calculate the highest probability.

$$H = \operatorname{argmax}_C \prod_{i=1}^n P(D_i|C) P(C) \quad (6)$$

3. K-Nearest Neighbors (KNN):

The k-NN algorithm is the most fundamental supervised machine learning algorithm used for classification. It classifies the given blood samples by using the majority of the classes in the clinical dataset's k-nearest neighbors. To find the nearest neighbors for a given data point, the algorithm typically employs the Euclidean distance metric. The distance metric formula is given in Eq. 7:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^n w_k (f_k(x_i) - f_k(x_j))^2} \quad (7)$$

where x = (f1, f2, f3, ..., fn), n is the number of attributes, fk is the kth attribute with its weight denoted by wk, and d (xi, xj) is the distance between xi and xj [46].

Pseudocode for the k-Nearest Neighbor Algorithm [46]

Input: D=[X,Y]; X,Y(class labels)

Output: Test case class

1. Initialize the model with random values for the weights, w.
2. Split the dataset, D, into two parts: a training set, T, and a testing set, T1.
3. Read the training dataset, T.
4. For i=1 to n, do
 Compute distance d(Ti,T1).
End for
5. Compute set1 containing indices for the k smallest distances, d(Ti,T1).
6. Return a majority label for {Yi where i ∈ I}.

4. Logistic Regression:

Logistic regression predicts using a logistic function. The logistic function is a sigmoid function that takes a real-value number as input and maps it between 0 and 1 [47]. The 15 features selected using the feature selection algorithm are given as input and the class membership probability is calculated using Eq. 8

$$y = \frac{e^{(b_0 + b_1 * x)}}{(1 + e^{(b_0 + b_1 * x)})} \quad (8)$$

where the predicted output is denoted by y, b0 is the intercept term, and b1 is the coefficient of input x[48]. The binary classification made is based on the value of y. Here the binary class value is either 0 or 1. A class value of 0 is COVID-19 negative and a class value of 1 is COVID-19 positive as shown in Eq. 9 and Eq. 10.

$$0 \text{ if } y < 0.5 \quad (9)$$

$$1 \text{ if } y \geq 0.5 \quad (10)$$

Pseudocode for Logistic Regression

Input: D = [X, Y]; X (dataset with n features), Y (class labels)

Output: Test case class

1. Initialize the model with random values for the weights, w.
2. Split the dataset, D, into two parts: a training set, T, and a testing set, T1.
3. Read the training dataset, T.
4. For each data in the training set, T,
 Calculate the probability using the formula in Eq. 8
5. If y<0.5, assign the class label 0, otherwise assign the class label 1.

IV. COVID-19 PREDICTION PROCEDURE

An early screening procedure for predicting COVID-19 is given in the form of pseudocode. The blood test values are given as input and the output class gives the information whether the person is affected by COVID-19 or not. The working of COVID-19 prediction is given as a pseudocode below:

Pseudocode for COVID-19 Prediction

Input: COVID-19 Clinical Dataset D with Y classes

Begin

D1: Convert categorical values to numerical values using one-hot encoding // data preprocessing

X: Select relevant features from D1 // feature selection

T: 80% samples from X // training set

T1: Remaining 20% of samples from X // testing set

N: Number of samples in T

F: Feature labels from f1 to fn

C: Classification algorithm

For each feature in f1 ... fn,

 Construct a new label vector for the Y classes.

Apply T to C // the classifier is trained using training dataset.

Testing data (T1) is given as input to the trained classifier.

Calculate the confusion matrix for T1.

Evaluate the performance of the classifier using the confusion matrix.

Choose the classifier and feature selection algorithm with the best accuracy.

End

Output: Prediction of COVID-19 using the best feature selection and classifier algorithm

V. EXPERIMENTAL FINDINGS AND DISCUSSION

This section included several experiments to determine the best feature selection and classification algorithm for COVID-19 prediction. The first experiment was carried out to identify the best feature selection algorithm which selects significant features from the dataset. The second experiment was carried out to find a suitable classifier for the selected features. The clinical dataset taken for the experimental set-up is described in the following section.

A. Dataset Description

The features of the proposed dataset are age, gender, D-Dimer, C-Reactive Protein (CRP), Lactate DeHydrogenase (LDH), total number of white blood cells (TC), platelet count (PC), packed cell volume (PCV), monocytes, eosinophils, basophils, mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), erythrocyte sedimentation rate (ESR (60 min)), lymphocytes, serum glutamic-oxaloacetic transaminase (SGOT), random blood sugar (RBS), billirubin T, direct billirubin, indirect billirubin, DC:Neutrophils, emoglobin (HB), red blood cells (RBC), red cell distribution width RDW-CV, urea and creatinine. The last feature 'class' is used to identify whether the person is affected by COVID-19 or not, with '1' and '0' indicating COVID-19 positive and negative, respectively. Based on the 'class' feature, the dataset is divided into the types shown in Table II.

TABLE II. STATISTICAL INFORMATION OF THE DATASET

2000 Blood Samples			
Positive		Negative	
1000		1000	
Female	Male	Female	Male
408	592	480	520

B. Ground Truth and Predicted Output Using the Existing Classifiers

The details of ten patients were given as input to the existing SVM, Naïve Bayes, k-NN, and logistic regression classifiers to predict whether the person is affected by COVID-19 or not. The results are tabulated in Table III, wherein the actual values correspond to the physician's diagnosis outcome, and the predicted values correspond to the target values predicted.

The "actual" values presented in the table signify the outcomes provided by a physician, aligning with the model's predictions. This verification in a real-world context serves to showcase the model's practicality and its relevance within a clinical setting.

1. Performance Metrics

The performance metrics used for selecting the best feature selection method and classification algorithm are discussed below.

TABLE III. GROUND TRUTH AND PREDICTED OUTPUT USING EXISTING CLASSIFIERS

Input										
Features	Patients									
	1	2	3	4	5	6	7	8	9	10
AGE	67	35	55	78	74	60	40	28	39	62
GENDER	F	M	F	M	M	F	F	M	M	M
HB	10.8	15.9	12.8	15.2	11.9	18.6	18.3	13.6	16.3	14.9
TC	7400	4500	9400	17200	18300	5000	6200	12300	20400	7400
DC: NEUTROPHILS	58	61	70	85	25	85	60	89	23	70
LYMPHOCYTES	23	31	21	15	20	12	38	17	12	27
EOSINOPHILS	2	5	1	0	1	0	0	2	1	1
MONOCYTES	5	6	2	2	1	1	4	2	2	1
BASOPHILS	1	1	1	0	1	1	1	0	0	1
ESR(60 MIN)	25	10	74	57	12	13	15	45	83	43
PC	2.8	1.9	4.3	2.6	1.6	4.4	1.3	2.2	4.5	2.5
PCV	33	48	31	39	36	27	47	37	38	38
MCV	83	89.8	91	88	89	74	86	89	87	85
MCH	30	27.9	31	32	32	24	30	31	31	31
MCHC	35	31.1	35	36	35	32	35	34	35	36
RBC	3.9	5.3	3.3	4.5	4	3.5	5.4	4	4.3	4.4
RDW-CV	15.2	15.5	15.3	15.1	11.8	12.4	12.5	11.5	15.1	18.6
RBS	169	189	196	230	220	169	143	350	467	220
UREA	15	35	37	41	58	22	20	24	60	30
CREATININE	0.8	1.5	1.3	1.2	1.2	1	0.8	1.2	1.3	1.1
CRP	79.8	10	40	60	55	43	3	36	26	15
D-DIMER	550	240	200	150	140	110	115	110	100	600
LDH	112	186	294	289	190	170	190	278	147	282
DIRECT BILLIRUBIN	0.4	0.3	0.3	0.4	0.4	0.6	0.3	0.4	0.3	0.3
BILLIRUBIN T	0.8	0.9	0.5	0.7	0.8	1.5	0.5	0.8	0.5	0.7
INDIRECT BILLIRUBIN	0.4	0.6	0.2	0.3	0.4	0.9	0.2	0.4	0.2	0.4
SGOT	29	25	27	20	35	54	30	39	38	37
Classifier	Output									
	Actual	1	0	1	1	1	1	0	1	1
SVM	Predicted	1	0	1	0	1	1	1	1	1
NB	Predicted	1	1	1	0	1	1	1	1	1
K-NN	Predicted	1	1	1	0	1	1	1	0	1
LR	Predicted	1	0	1	0	1	1	1	0	1

The performance of the classifier can be illustrated using confusion matrix. Most of the metrics are measured using confusion matrix. The accuracy, recall, precision, F1-score, and AUC were used to evaluate the results. Table IV displays the confusion matrix for COVID-19 prediction.

TABLE IV. CONFUSION MATRIX FOR COVID-19 PREDICTION

	Has COVID-19 Disease	Does not have COVID-19 Disease
Has COVID-19 Disease	True positive	False Positive
Does not have COVID-19 Disease	False Negative	True Negative

True Positive (TP) :

If the actual class is COVID-19 positive and the model also predicts the class value as COVID-19 positive, then it is termed as True Positive.

True Negative (TN):

If the actual class is COVID-19 negative and predicted class value is also COVID-19 negative, then it is termed as True Negative.

False Positive (FP):

If the actual class is COVID-19 positive but the predicted class result is COVID-19 negative, then it is termed as False Positive

False Negative (FN):

If the actual class is COVID-19 negative but the predicted class result is COVID-19 positive, then it is termed as False Negative.

Accuracy:

Accuracy is computed by adding the number of correctly predicted positive and negative predictions and then dividing it by all types of predictions (TP, TN, FP, FN) [49] as shown in Eq. 11.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (11)$$

Precision:

Precision is the fraction of number of correctly predicted positive instances and the total number of correct or incorrect predicted positive instances (TP, FP) [50]. Precision is also termed the Positive Predictive Rate (PPR) as shown in Eq. 12.

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (12)$$

Recall:

Recall is the fraction of correctly predicted positive (TP) instances and the sum of correctly predicted positive and incorrectly predicted negative instances (TP, FN) [50] as shown in Eq.13. It is otherwise called the True Positive Rate (TPR).

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (13)$$

F1-score:

F1-score is calculated as the weighted average of precision and recall as shown in Eq. 14. Since it takes into account false positive and false negative predictions, these metric measures accuracy for uneven datasets better [50].

$$\text{F1-Score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (14)$$

AUC:

AUC stands for Area under the curve. It is the measure of how well it distinguishes between each class. It is also known as Receiver Operator Characteristics (ROC) curve summary. It is used as a metric in binary classification problem.

C. Validation Methods

This section deals with the two types of cross-fold and split dataset validation methods used in the research.

K-fold validation:

The K-fold validation method trains and evaluates the model “k” times for different samples [51]. Performance metrics are used to evaluate each fold, and the fold with the highest accuracy is selected the best.

Data split validation:

Based on the number of samples, the dataset is divided into a train set split and a test set split. The split ratio normally commences with 80:20, 75:25, and 70:30, and goes on likewise. Metrics are used in every split to measure the performance of the model and the split with the best accuracy.

D. Comparison of State-of-the Art Benchmarks in COVID-19 Prediction

In this section, a table is presented to outline the different machine learning algorithms employed in preprocessing, feature selection, and classification techniques. Table V compiles the state-of-the-art benchmarks in predicting COVID-19.

TABLE V. COMPARISON OF STATE-OF THE ART BENCHMARKS IN COVID-19 PREDICTION

Ref. No.	Dataset Used	Pre Processing	Feature Extraction / Selection	Classification/ Techniques Used	Accuracy (%)
		Noise removal / Handling Outliers			
[15]	IRCCS Ospedale San Raffaele	N/A	Feature Importance	Random Forest	82
[17]	Albert Einstein Hospital dataset	Highly correlated attributes are eliminated to reduce noise in the data.	Pearson Cor-relation and feature importance	Logistic regression, random forest, k nearest neighbours and Xgboost	92
[21]	Albert Einstein Hospital in São Paulo, Brazil	N/A	Correlation Matrix and The Chi-Squared Test	Ensemble of Support Vector Machines, Adaptive Boosting, Random Forest and K-Nearest Neighbors	69.9
[23]	Albert Einstein Hospital (Kaggle)	N/A	N/A	K Nearest Neighbor, Radial Basis Function, Naive Bayes, kStar, PART, Random Forest, Decision Tree, OneR, Support Vector Machine and Multi-Layer Perceptron	88
	Proposed Clinical Dataset	iForest to Handle Outliers	Mutual Information	SVM	94.8

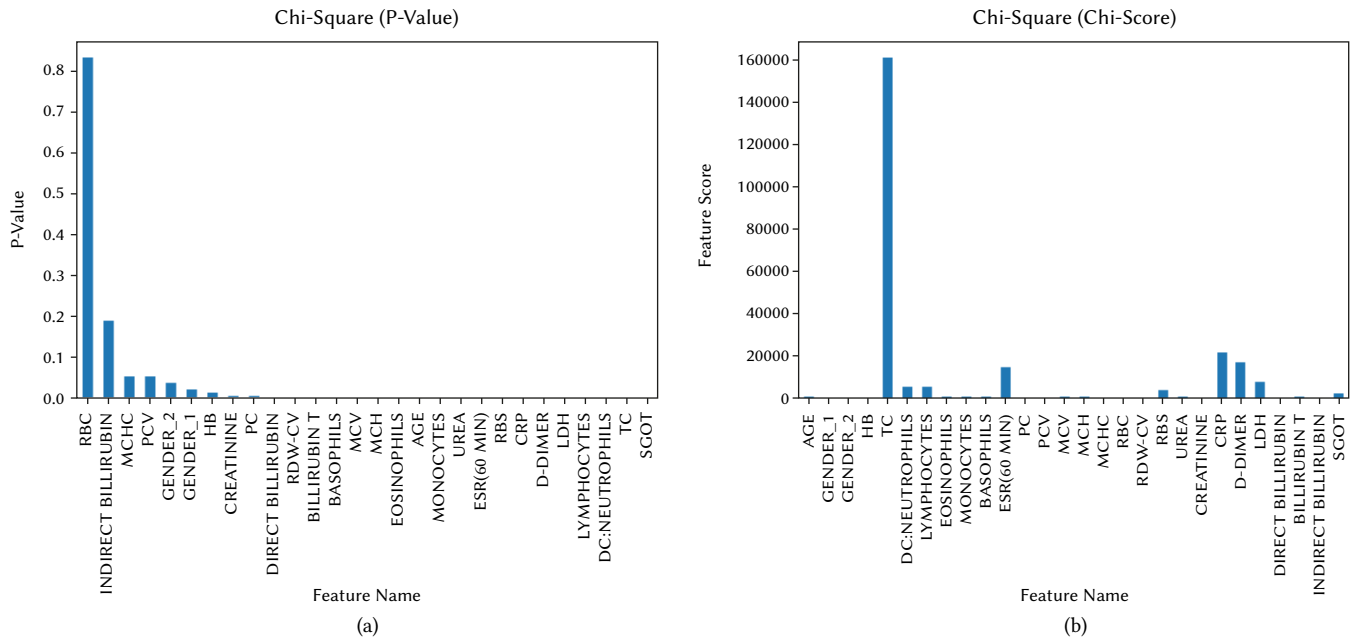


Fig. 4. (a) P-Value of the features using Chi-Square. (b) Chi-Score of the features using Chi-Square.

Based on the information provided in the table above, it is clear that the model using Albert Einstein Hospital (Brazil) dataset and the proposed clinical dataset that checks and handle outliers demonstrates improved accuracy than other models. Therefore, addressing outliers has effectively enhanced the model's robustness.

E. Finding the Best Features Using Feature Selection Methods

In the feature selection stage, which is indispensable to designing the model, the most appropriate features that maximize the model's performance are chosen. This section shows the results of the feature selection algorithm in the process of selecting the features from the clinical dataset. Feature selection methods used for the study include the ANOVA-F, chi-square, mutual information and Pearson correlation filter-based methods, as well as the RFE and SFS wrapper-based methods.

1. Chi-Square Test:

The outputs of the chi-squared test are the p-value and the chi score. A large p-value shows target-independent input features that are not selected for training. Target-dependent features with a high chi score, on the other hand, are selected for training. Fig. 4(a) and 4(b) show the graph representing the chi-square, based on the p-values and chi-score, respectively. The threshold for the chi-score is set to 100 and for the p-value to 0.05. Features selected for training include the MCV, MCH, Eosinophils, age, monocytes, urea, ESR (60 min), RBS, CRP, D-Dimer, LDH, lymphocytes, DC: Neutrophils, TC, and SGOT. These features have a p-value and a chi-score that are less than and greater than the threshold, respectively.

2. ANOVA-F:

In the ANOVA-F, the impact of the feature with the target variable is determined by the feature's variance. A low score implies that the feature has no impact on the target feature. Fig. 5. graphically depicts the feature score of all the features in the dataset using ANOVA-F feature selection. The top 15 features selected for classification are the TC, monocytes, RBS, Direct Billirubin, DC:Neutrophils, lymphocytes, basophils, ESR (60 min), MCH, D-Dimer, CRP, LDH, Eosinophils, Billirubin T and SGOT.

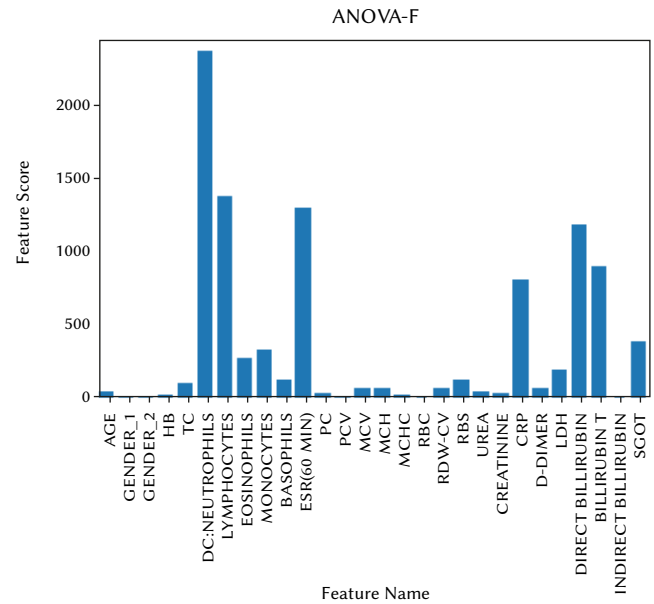


Fig. 5. Feature and its score using ANOVA-F.

3. Mutual Information:

Mutual information calculates the information gain for each feature. Fig. 6. shows the feature score calculated using mutual information for each feature. The top 15 features with high information gain are selected for classification. The selected features are DC: Neutrophils, Lymphocytes, CRP, Billirubin T, ESR (60 min), Direct Billirubin, D-Dimer, LDH, MCV, MCH, RBS, RBC, UREA, Eosinophils, and PC.

4. Pearson Correlation:

Fig. 7. Depicts the correlation matrix of various clinical dataset features. Highly correlated features are selected for classification. The selected features include DC:Neutrophils, Eosinophils, ESR (60 min), monocytes, basophils, MCV, MCH, RBS, Direct Billirubin, CRP, RDW-CV, Billirubin T, LDH, SGOT, and D-Dimer.

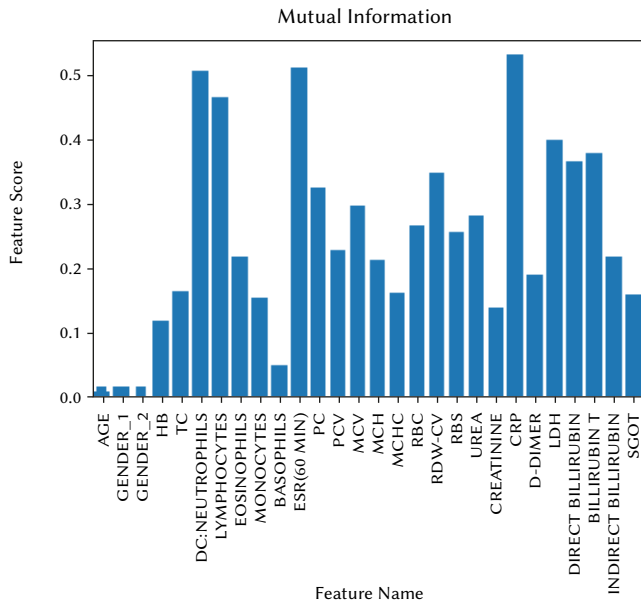


Fig. 6. Feature and its score using Mutual Information.

5. Recursive Feature Elimination (RFE):

Table VI shows a list of features selected using Recursive Feature Elimination (RFE). The selected features are those with a 'true' value in the RFE_Support column, while the RFE_Ranking column provides information on the rank of each feature. Features with '1' in the RFE_Ranking and 'True' in the RFE_Support are selected as the best which include Lymphocytes, DC: Neutrophils, Eosinophils, Monocytes, Basophils, ESR (60 min), PC, PCV, MCV, RBS, RDW-CV, urea, creatinine, CRP, D-Dimer and LDH.

TABLE VI. FEATURES SELECTED USING RFE ALGORITHM

S. No.	Feature Name	RFE_Support	RFE_Ranking
1	AGE	False	14
2	GENDER_1	False	13
3	GENDER_2	False	12
4	HB	False	9
5	TC	False	8
6	DC: NEUTROPHILS	True	1
7	LYMPHOCYTES	True	1
8	EOSINOPHILS	False	2
9	MONOCYTES	True	1
10	BASOPHILS	True	1
11	ESR (60 MIN)	True	1
12	PC	True	1
13	PCV	True	1
14	MCV	True	1
15	MCH	False	10
16	MCHC	False	11
17	RBC	False	6
18	RDW-CV	True	1
19	RBS	True	1
20	UREA	True	1
21	CREATININE	True	1
22	CRP	True	1
23	D-DIMER	True	1
24	LDH	True	1
25	DIRECT BILLIRUBIN	False	3
26	BILLIRUBIN T	False	4
27	INDIRECT BILLIRUBIN	False	5
28	SGOT	False	7

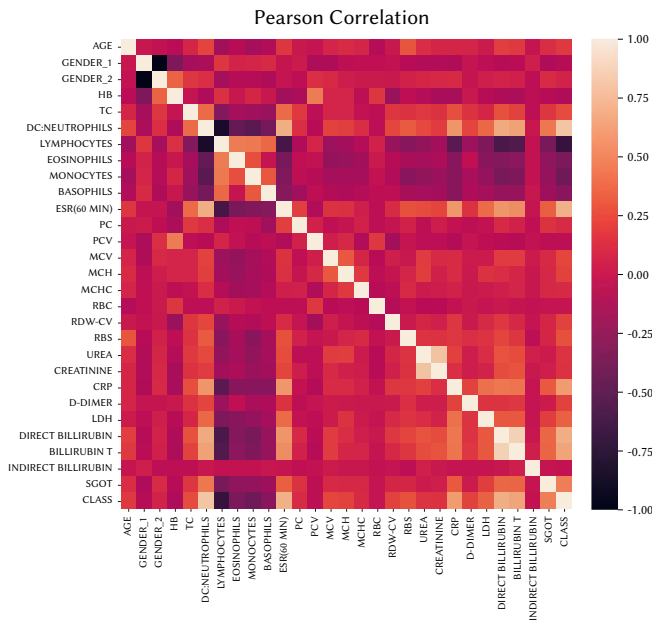


Fig. 7. Correlation Matrix of Clinical Dataset features.

6. Sequential Forward Selection (SFS):

Table VII shows a list of features and the average score of the feature subset selected using the Sequential Forward Selection (SFS) algorithm. The Feature_Names column displays a list of features selected in each iteration and the Avg_Score column gives the average

of the feature score of the selected features in each iteration. Features selected by the SFS algorithm are Gender_1, Gender_2, HB, TC, DC Neutrophils, Lymphocytes, Eosinophils, Monocytes, Basophils, ESR (60 min), PC, PCV, MCV, CRP, and LDH.

F. Features Selected by Various Feature Selection Method:

The experiments above have selected certain features by different feature selection method. Some features are selected by more than one feature selection method. Table VIII presents an analysis of the votes gained by each feature.

It is found from the above table, features like CRP, DC neutrophils, lymphocytes, eosinophils, basophils, ESR (60 min), MCV, RBS, D-dimer, LDH, direct bilirubin, and bilirubin T have a high vote.

G. Comparison of Classification Techniques Performance Based on Feature Selection Methods

Table VIII shows the comparison of classification technique performance based on various feature selection techniques. The dataset contains 28 features, of which 15 were selected using the ANOVA-F, chi-square test, mutual information, Pearson correlation, RFE and SFS feature selection methods. These 15 features are used for classifying patients' with COVID-19. The efficiency of the classifier is determined by various performance metrics. The confusion matrix helps to view the efficiency of the classifier pictorially. Fig.8, 9, 10, and 11 show, respectively, the confusion matrix obtained for the test data after training the model using the SVM, Naïve Bayes, k-NN and logistic regression classifiers. The classifiers work with the features selected using mutual information.

TABLE VIII. FEATURES SELECTION METHOD AND SELECTED FEATURES

Name of the Feature	List of Selected Features with Feature Selection (15)					
	Filter Method				Wrapper Method	
	ANOVA-F	Chi-Square	Mutual Information	Pearson Correlation	RFE	SFS
AGE						
GENDER_1						✓
GENDER_2						✓
HB						✓
TC	✓	✓				✓
DC: NEUTROPHILS	✓	✓	✓	✓	✓	✓
LYMPHOCYTES	✓	✓	✓		✓	✓
EOSINOPHILS	✓	✓	✓	✓		✓
MONOCYTES	✓	✓		✓	✓	✓
BASOPHILS	✓			✓	✓	✓
ESR(60 MIN)	✓	✓	✓	✓	✓	✓
PC			✓		✓	✓
PCV					✓	✓
MCV		✓	✓	✓	✓	✓
MCH	✓	✓	✓	✓		
MCHC						
RBC			✓			
RDW-CV				✓	✓	
RBS	✓	✓	✓	✓	✓	
UREA		✓	✓		✓	
CREATININE					✓	
CRP	✓	✓	✓	✓	✓	✓
D-DIMER	✓	✓	✓	✓	✓	
LDH	✓	✓	✓	✓	✓	✓
DIRECT BILLIRUBIN	✓		✓	✓		
BILLIRUBIN T	✓		✓	✓		
INDIRECT BILLIRUBIN						
SGOT	✓	✓		✓		

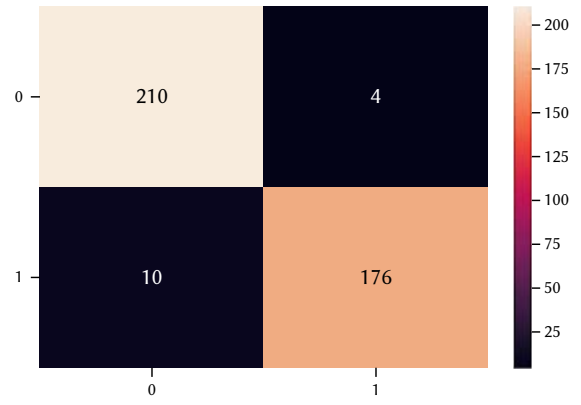


Fig. 8. Confusion Matrix – SVM.

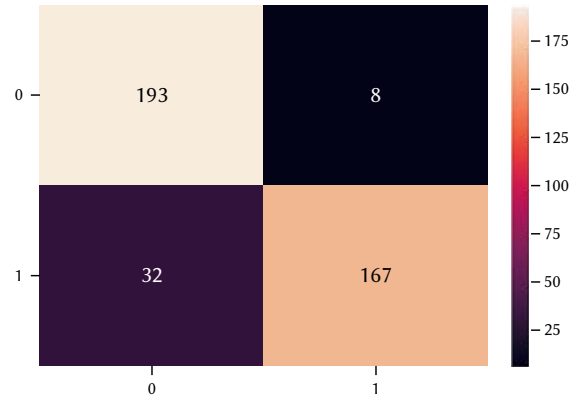


Fig. 9. Confusion Matrix –Naïve Bayes.

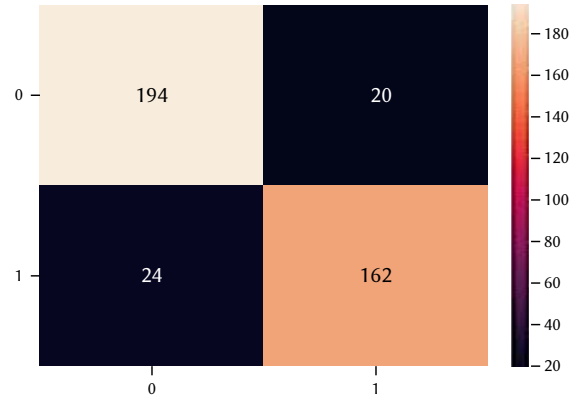


Fig. 10. Confusion Matrix – k-NN.

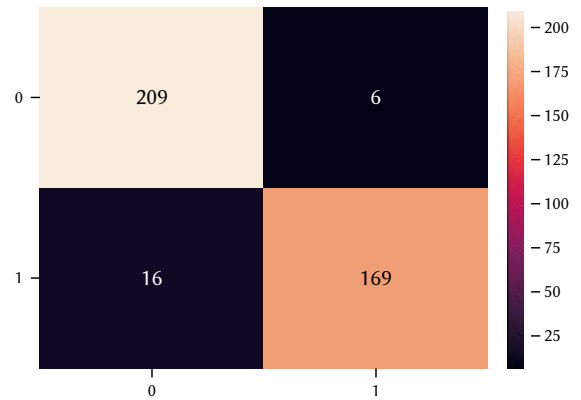


Fig. 11. Confusion Matrix – LG.

Note : 1 - COVID-19 Positive ; 0 – COVID-19 Negative

TABLE VII. FEATURES SELECTED USING SFS ALGORITHM

S. No	Feature_Names	Avg_Score
1	CRP	0.71
2	CRP, LDH	0.75
3	TC, CRP, LDH	0.76
4	GENDER_1, TC, CRP, LDH	0.73
5	GENDER_1, GENDER_2, TC, CRP, LDH	0.76
6	GENDER_1, GENDER_2, HB, TC, CRP, LDH	0.78
7	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, CRP, LDH	0.74
8	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, CRP, LDH	0.75
9	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, CRP, LDH	0.79
10	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, CRP, LDH	0.78
11	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, CRP, LDH	0.81
12	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), CRP, LDH	0.83
13	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), PC, CRP, LDH	0.84
14	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, CRP, LDH	0.86
15	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, CRP, LDH	0.88
16	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, CRP, LDH	0.84
17	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, CRP, LDH	0.81
18	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, CRP, LDH	0.82
19	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, RDW-CV, CRP, LDH	0.76
20	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, RDW-CV, RBS, CRP, LDH	0.74
21	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, RDW-CV, RBS, UREA, CRP, LDH	0.82
22	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, RDW-CV, RBS, UREA, CFREATININE, CRP, LDH	0.78
23	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, RDW-CV, RBS, UREA, CFREATININE, CRP, D-DIMER, LDH	0.79
24	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, RDW-CV, RBS, UREA, CFREATININE, CRP, D-DIMER, LDH, DIRECT BILLIRUBIN	0.81
25	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, RDW-CV, RBS, UREA, CFREATININE, CRP, D-DIMER, LDH, DIRECT BILLIRUBIN, BILLIRUBIN T	0.79
26	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, RDW-CV, RBS, UREA, CFREATININE, CRP, D-DIMER, LDH, DIRECT BILLIRUBIN, BILLIRUBIN T, INDIRECT BILLIRUBIN, SGOT	0.77
27	AGE, GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, RDW-CV, RBS, UREA, CFREATININE, CRP, D-DIMER, LDH, DIRECT BILLIRUBIN, BILLIRUBIN T, INDIRECT BILLIRUBIN, SGOT	0.77

TABLE IX. PERFORMANCE OF DIFFERENT CLASSIFIERS WITH AND WITHOUT DIFFERENT FEATURE SELECTION METHODS

Feature Selection Algorithm	No. of Selected attribute	Classifiers	Performance Metrics				
			Accuracy (%)	Precision	Recall	F1-score	AUC
Without Feature selection	28	SVM	92.7	0.93	0.93	0.93	0.93
		Naïve Bayes	89.7	0.90	0.90	0.90	0.90
		KNN	69.7	0.75	0.70	0.68	0.70
		LR	91.7	0.92	0.92	0.92	0.93
Filter	ANOVA-F	15	SVM	93.7	0.94	0.94	0.94
			Naïve Bayes	92.7	0.92	0.92	0.93
			KNN	90	0.89	0.90	0.90
			LR	93.7	0.94	0.94	0.94
	Chi-Square	15	SVM	93.7	0.94	0.94	0.94
			Naïve Bayes	90.6	0.90	0.90	0.91
			KNN	89	0.89	0.89	0.89
			LR	91.25	0.92	0.91	0.92
	Mutual Information	15	SVM	94.8	0.95	0.95	0.95
			Naïve Bayes	90	0.91	0.90	0.90
			KNN	89	0.89	0.89	0.89
			LR	91.25	0.91	0.92	0.92
	Pearson Correlation	15	SVM	92.7	0.93	0.93	0.93
			Naïve Bayes	92.7	0.93	0.93	0.93
			KNN	88.5	0.87	0.89	0.89
			LR	92.7	0.93	0.93	0.93
Wrapper	RFE	15	SVM	91.6	0.92	0.92	0.92
			Naïve Bayes	88.5	0.89	0.89	0.89
			KNN	88.5	0.89	0.89	0.89
			LR	91.6	0.92	0.92	0.92
	SFS	15	SVM	92.7	0.93	0.93	0.93
			Naïve Bayes	92.7	0.92	0.93	0.93
			KNN	88.5	0.87	0.89	0.89
			LR	92.7	0.93	0.93	0.93

It is inferred from Table IX that the features selected by mutual information perform the best with the SVM classifier compared to other methods, producing 94.8% accuracy

H. Performance Evaluation of Feature Selection Techniques Using K-Fold Validation

According to the results, the SVM classifier paired with feature selection technique works well. Fold validation and training-testing data split validation helps to improve the efficiency of the model by providing us the best fold and split. Table IX shows the comparison of various feature selection techniques performance using the SVM

TABLE X. COMPARISON OF FEATURE SELECTION METHODS PERFORMANCE WITH SVM CLASSIFIER BASED ON FOLD VALIDATION

Metrics	Feature Selection Algorithm	Comparison of Feature Selection Methods Performance Based on Fold Validation				
		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Accuracy (%)	ANOVA-F	61.6	65.04	76.5	88.62	89.3
	Chi-Square	72.4	71.04	81.9	90.6	91.7
	Mutual Information	94.1	93.9	94.08	93.9	94.5
	Pearson Correlation	92.1	91.9	92.08	91.9	92.4
	RFE	89.1	88.8	89.1	88.8	89.8
	SFS	91.5	91.4	91.5	91.5	91.9
Precision	ANOVA-F	0.62	0.65	0.77	0.89	0.90
	Chi-Square	0.73	0.71	0.82	0.91	0.92
	Mutual Information	0.93	0.92	0.94	0.94	0.95
	Pearson Correlation	0.91	0.92	0.91	0.91	0.92
	RFE	0.89	0.88	0.89	0.88	0.89
	SFS	0.91	0.90	0.91	0.91	0.91
Recall	ANOVA-F	0.62	0.65	0.77	0.89	0.89
	Chi-Square	0.73	0.71	0.82	0.91	0.92
	Mutual Information	0.93	0.93	0.94	0.94	0.95
	Pearson Correlation	0.91	0.91	0.91	0.91	0.92
	RFE	0.89	0.88	0.89	0.88	0.89
	SFS	0.91	0.90	0.91	0.91	0.91
F1-Score	ANOVA-F	0.62	0.65	0.77	0.89	0.89
	Chi-Square	0.73	0.71	0.82	0.91	0.91
	Mutual Information	0.93	0.92	0.94	0.94	0.95
	Pearson Correlation	0.91	0.91	0.91	0.91	0.92
	RFE	0.89	0.88	0.89	0.88	0.89
	SFS	0.91	0.90	0.91	0.91	0.91

classifier in COVID-19 prediction. To find the effective fold for all filter and wrapper based feature selection methods, cross-fold validation is used. This experiment divides the dataset into 5 fold ranging from 1 to 5 and the above mentioned performance metrics are used for evaluation. Table X shows the comparison of various feature selection technique performance with the SVM classifier.

It is observed, from the results of Table X that 5th fold gives the best results. Moreover, the performance metrics show that the mutual information technique outperforms all the others.

I. Performance Evaluation of Feature Selection Techniques Using Data Splitting Validation

Many researchers do not focus on fold or split validation. The importance of data splitting is highlighted in this research, with experiments carried out to determine the suitable split for testing and training. The above mentioned metrics are used to evaluate the performance of feature selection methods with SVM classifier in order to determine the best fold and data splitting range for predicting

TABLE XI. COMPARISON OF THE PERFORMANCE OF FEATURE SELECTION METHODS BASED ON DATA SPLITTING VALIDATION

Metrics	Feature Selection Algorithms	Comparison of the performance of feature selection methods based on data splitting validation												
		20-80	25-75	30-70	35-65	40-60	45-55	50-50	55-45	60-40	65-35	70-30	75-25	80-20
Accuracy (%)	ANOVA-F	80.6	80.6	80.2	81.2	81.9	80.5	82.1	81.6	83.1	84.9	85.8	89.1	93.7
	Chi-Square	80.1	81.5	82.7	83.6	84.2	85.6	86.1	87.7	88.8	89.9	90.1	91.7	93.7
	Mutual Information	83.2	84.9	85.2	86.8	87.2	88.8	89.4	90.2	91.8	92.6	93.1	93.8	94.8
	Pearson Correlation	80.6	81.7	80.3	82.3	83.4	85.1	86.7	87.1	88.4	89.1	90.5	91.2	92.7
	RFE	78.5	79.2	80.6	81.1	82.1	83.4	84.6	93.5	85.1	86.9	88.1	90.1	91.6
	SFS	75.2	76.2	77.8	78.2	80.2	81.5	82.9	84.1	88.2	89.4	90.1	91.8	92.7
Precision	ANOVA-F	0.78	0.80	0.80	0.81	0.82	0.81	0.82	0.82	0.83	0.85	0.86	0.89	0.94
	Chi-Square	0.80	0.81	0.83	0.84	0.84	0.86	0.86	0.88	0.89	0.90	0.91	0.92	0.94
	Mutual Information	0.84	0.85	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.93	0.93	0.94	0.95
	Pearson Correlation	0.81	0.82	0.81	0.82	0.83	0.85	0.87	0.87	0.88	0.89	0.91	0.92	0.93
	RFE	0.78	0.79	0.81	0.81	0.82	0.82	0.85	0.84	0.85	0.87	0.88	0.90	0.92
	SFS	0.75	0.77	0.78	0.78	0.80	0.82	0.83	0.84	0.88	0.89	0.90	0.91	0.93
Recall	ANOVA-F	0.77	0.81	0.81	0.82	0.82	0.81	0.82	0.82	0.83	0.85	0.86	0.89	0.94
	Chi-Square	0.79	0.81	0.83	0.84	0.83	0.86	0.86	0.87	0.88	0.89	0.90	0.92	0.94
	Mutual Information	0.85	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.93	0.94	0.95	0.94	0.95
	Pearson Correlation	0.81	0.82	0.81	0.82	0.83	0.85	0.87	0.87	0.88	0.89	0.91	0.92	0.93
	RFE	0.78	0.79	0.81	0.81	0.82	0.82	0.85	0.84	0.85	0.87	0.88	0.90	0.92
	SFS	0.75	0.77	0.78	0.78	0.80	0.82	0.83	0.84	0.88	0.89	0.90	0.91	0.93
F1 Score	ANOVA-F	0.77	0.81	0.81	0.82	0.82	0.81	0.82	0.82	0.83	0.85	0.86	0.89	0.94
	Chi-Square	0.79	0.81	0.83	0.81	0.83	0.86	0.84	0.87	0.88	0.89	0.89	0.91	0.94
	Mutual Information	0.85	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.93	0.94	0.95	0.94	0.95
	Pearson Correlation	0.81	0.82	0.81	0.82	0.83	0.85	0.87	0.87	0.88	0.89	0.91	0.92	0.93
	RFE	0.78	0.79	0.81	0.81	0.82	0.82	0.85	0.84	0.85	0.87	0.88	0.90	0.92
	SFS	0.75	0.77	0.78	0.78	0.80	0.82	0.83	0.84	0.88	0.89	0.90	0.91	0.93

TABLE XII. COMPARISON OF OTHER DATASET PERFORMANCE WITH SVM CLASSIFIER

S. No.	Ref. No.	Dataset	No. of Features	Instances	Features Selected	Accuracy (%)
1.	[15]	IRCCS Ospedale San Raffaele	15	219 (COVID-19 positive - 177 COVID-19 Negative - 102)	Gender, Age, WBC, platelets, CRP, AST, ALT, GGT, LDH, Neutrophils, Lymphocytes, Monocytes, Eosinophils, Basophils, Swab	82
2.	[21]	Albert Einstein Hospital in São Paulo, Brazil	111	5644 (COVID-19 positive - 558 COVID-19 Negative - 5086)	Monocytes, Age, Red Blood Cells, Serum Glucose Hematocrit, Hemoglobin, Leukocytes, Lymphocytes, Mean Platelet Volume, Creatinine, Calcium, Magnesium, Potassium, Sodium, Urea, Vitamin B12, Phosphor	69.79
3.	[23]	Albert Einstein Hospital (Kaggle)	72	1624 (COVID-19 positive - 786 COVID-19 Negative - 838)	LDH, AST, FG, CA, PCR, GLU, ALT, CO2POC, SO2POC, GLUEMO, WBC, FCOPOC, RDW, HHBPOC, AGE, HCT, FO2POC, BAT, XDP, GGT.	88
4.	-	Proposed COVID-19 Clinical Dataset	27	2000 (COVID-19 positive - 1000 COVID-19 Negative - 1000)	DC:Neutrophils, Lymphocytes, CRP, Billirubin T, ESR(60 Min), Direct Billirubin, D-Dimer, LDH, MCV, MCH, RBS, RBC, UREA, Indirect Billirubin, PC	94.8

COVID-19 using clinical data. Table XI lists a comparison of the performance of feature selection methods with the SVM classifier to predict COVID-19. To get the best training and testing splitting range, the split is listed in ranges from 20 - 80% to 80% - 20% as depicted in Table XI.

It is evident from Table XI that the 80%-20% training-testing data splitting shows high accuracy. The result shows that mutual information outperforms other feature selection techniques.

J. Comparing the Performance of the Datasets

This section compares the performance of open source clinical datasets with the proposed clinical dataset. It is found from the literature survey that open source datasets are the preferred choice for model-building. The total number of features and instances, as well as features chosen by the feature selection algorithm, are analysed. The features were classified using the SVM classifier and its performance.

It is observed from Table XII that the open source datasets used have imbalanced data, unlike the proposed dataset. This proposed dataset has been used to build a model that selects relevant features and predicts COVID-19 using the SVM classifier with 94.8% accuracy, outclassing other datasets. The hyperparameters such as C (penalty parameter), kernel, gamma, coef0 can be tuned to improve the performance of the model.

VI. CONCLUSION

This research was carried out to publish a new clinical dataset on GitHub. Further, it focused on selecting the best features using feature selection techniques and finding a suitable classifier to predict COVID-19. To this end, a literature survey was completed to examine the feature selection methods and classification algorithms used for COVID-19 prediction using a clinical dataset. Different experiments were conducted using the clinical dataset in order to determine the suitable feature selection algorithm that selects the most relevant features, along with an appropriate classifier for the prediction. Based on the experiments, the mutual information filter-based feature selection algorithm was identified to be the best of its kind. The SVM classifier, with a high 94.8% accuracy, outperformed the rest. While the model excels at predicting COVID-19, its primary limitation lies in its lack of generalizability, stemming from its reliance on data from a single hospital for model training. Future directions include extending the research by modifying the mutual information algorithm to select the best feature to enhance the performance of the classifier. Likewise, two classifiers can be combined to form an ensemble classifier that can be used to build a high-performance classifier for COVID-19 prediction using the clinical dataset.

REFERENCES

- [1] T. Singhal, "A Review of Coronavirus Disease-2019 (COVID-19)," *Indian Journal of Pediatrics*, vol. 87, no.4, pp.281-286, 2020, doi: 10.1007/s12098-020-03263-6.
- [2] S. Mubareka, J. B. Gubbay, W. C. Chan, "Diagnosing COVID-19: the disease and tools for detection," *American Chemical Society Nano*, vol. 14 no.4, pp. 3822-35, 2020, doi:10.1021/acsnano.0c02624.
- [3] D. Li, D. Wang, J. Dong, N. Wang, H. Huang, H. Xu, C. Xia, "False-negative results of real-time reverse-transcriptase polymerase chain reaction for severe acute respiratory syndrome coronavirus 2: role of deep-learning-based CT diagnosis and insights from two cases," *Korean journal of radiology*, vol. 21, no. 4, pp. 505-508, 2020, doi: 10.3348/kjr.2020.0146.
- [4] A. Ulhaq, J. Born, A. Khan, D. P. S. Gomes, S. Chakraborty, M. Paul, "COVID-19 Control by Computer Vision Approaches: A Survey," *IEEE Access*, vol. 8, pp. 179437-179456, 2020, doi: 10.1109/ACCESS.2020.3027685.
- [5] J. Bao, C. Li, K. Zhang, H. Kang, W.Chen, B. Gu, "Comparative analysis of laboratory indexes of severe and non-severe patients infected with COVID-19," *Clinica Chimica Acta*, vol. 509, pp. 180-194, 2020, doi: 10.1016/j.cca.2020.06.009.
- [6] B. E. Fan, "Hematologic parameters in patients with COVID-19 infection: a reply," *American journal of hematology* vol. 95, no. 6, 2020, doi: 10.1002/ajh.25774.
- [7] Y. Gao, T. Li, M. Han, X. Li, D. Wu, Y. Xu, Y. Zhu, Y. Liu, X. Wang, L. Wang, "Diagnostic utility of clinical laboratory data determinations for patients with the severe COVID-19," *Journal of medical virology*, vol. 92, no. 7, pp. 791-796, 2020, doi: 10.1002/jmv.25770.
- [8] T. A. Khartabil, H. Russcher, A. Ven, Y. B. Rijke, "A summary of the diagnostic and prognostic value of hemocytometry markers in COVID-19 patients," *Critical reviews in clinical laboratory sciences*, vol. 57, no. 6, pp. 415-431, 2020, doi:10.1080/10408363.2020.1774736.
- [9] A. J. Rodriguez-Morales, J. A. Cardona-Ospina, E. Gutiérrez-Ocampo, R.Villamizar-Peña, Y. Holguin-Rivera, J. P. Escalera-Antezana, et al., "Clinical, laboratory and imaging features of COVID-19: A systematic review and meta-analysis," *Travel medicine and infectious disease*, vol.34, 2020, doi: 10.1016/j.tmaid.2020.101623.
- [10] J. A. Siordia, "Epidemiology and clinical features of COVID-19: A review of current literature," *Journal of Clinical Virology*, vol. 127, 2020, doi: 10.1016/j.jcv.2020.104357.
- [11] Y. Liu, Y. Yang, C. Zhang, F.Huang, F. Wang, J. Yuan, et al., "Clinical and biochemical indexes from 2019-nCoV infected patients linked to viral loads and lung injury," *Science China Life Sciences*, vol. 63, no. 3, pp. 364-74, 2020, doi: 10.1007/s11427-020-1643-8.
- [12] L. Muhammad, M. M. Islam, S. S. Usman, S. I. Ayon, "Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery," *SN Computer Science*, vol. 1, no. 4, pp. 1-7, 2020, doi: 10.1007/s42979-020-00216-w.
- [13] J. Wu, P. Zhang, L. Zhang, W. Meng, J. Li, C. Tong, Y. Li, J. Cai, Z. Yang, J. Zhu et al., "Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results," *MedRxiv*, 2020, doi:10.1101/2020.04.02.20051136.
- [14] A. Bastug, H. Bodur, S. Erdogan, D. Gokcinar, S. Kazancioglu, B. D. Kosovali, B. O. Ozbay, G. Gok, I. O. Turan, G. Yilmaz, C. C. Gonen, F. M. Yilmaz, "Clinical and laboratory features of COVID-19: Predictors of severe prognosis," *International Immunopharmacology*, vol. 88, no. 106950, 2020, doi: 10.1016/j.intimp.2020.106950.
- [15] D. Brinati, A. Campagner, D. Ferrari, M. Locatelli, G. Banfi, and F. Cabitza, "Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study," *Journal of medical systems*, vol. 44, no. 8, pp. 1-12, 2020, doi:10.1007/s10916-020-01597-4.
- [16] M. Kukar, G. Gunčar, T. Vovko et al., "COVID-19 diagnosis by routine blood tests using machine learning," *Scientific Reports*, vol. 11, no. 10738, 2021, doi:10.1038/s41598-021-90265-9.
- [17] K. Chadaga, S. Prabhu, K. V. Bhat, S. Umakanth and N. Sampathila., "Medical diagnosis of COVID-19 using blood tests and machine learning," *Journal of Physics: Conference Series*, vol. 2161(1), 2022, doi:10.1088/1742-6596/2161/1/012017
- [18] M. AlJame, I. Ahmad, A. Imtiaz, A. Mohammed, "Ensemble learning model for diagnosing COVID-19 from routine blood tests," *Informatics in Medicine Unlocked*, vol. 1, no. 21, 2020, doi: 10.1016/j.imu.2020.100449
- [19] A. F. M. Batista, J.L. Miraglia, T. H. R. Donato, A. D. P. C. Filho, "COVID-19 diagnosis prediction in emergency care patients: A machine learning approach," *medRxiv*, 2020, doi: 10.1101/2020.04.04.20052092 [CrossRef].
- [20] V. A. F. Barbosa, J. C. Gomes, M. A. Santana, J. E. A. Albuquerque, R. F. Souza, R. E. Souza, W. P. Santos, "Heg.IA: an intelligent system to support diagnosis of COVID-19 based on blood tests," *Research on Biomedical Engineering*, vol. 38, no. 1, pp. 99-116, 2022, doi: 10.1007/s42600-020-00112-5.
- [21] M. Almansoor and N. M. Hewahi, "Exploring The Relation Between Blood Tests And Covid-19 Using Machine Learning," *International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, pp. 1-6, 2020, doi: 10.1109/ICDABI51230.2020.9325673.
- [22] F. Cabitza, A. Campagner, D. Ferrari, C. Di Resta, D. Ceriotti, E. Sabetta, A. Colombini, E. De Vecchi, G. Banfi, M. Locatelli, A. Carobene, "Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests," *Clinical Chemistry and Laboratory Medicine*, vol. 59, no. 2, pp. 421-431, 2021, doi: 10.1515/cclm-2020-1294.
- [23] A. Akhtar, S. Akhtar, B. Bakhtawar, A.A. Kashif, N. Aziz, M. S. Javeid, "COVID-19 Detection from CBC using Machine Learning Techniques," *International Journal of Technology, Innovation and Management*, vol. 1, no. 2, 2021, doi:10.54489/ijtim.v1i2.22.
- [24] O. O. Abayomi-Alli, R. Damaševičius, R. Maskeliūnas, S. Misra, "An Ensemble Learning Model for COVID-19 Detection from Blood Test Samples," *Sensors (Basel)*, vol. 22, no. 6, 2022, doi: 10.3390/s22062224.
- [25] H. Gong, M. Wang, H. Zhang, M.F. Elahe, M. Jin, "An Explainable AI Approach for the Rapid Diagnosis of COVID-19 Using Ensemble Learning Algorithms," *Front Public Health*, vol. 10, no. 874455, 2022, doi:10.3389/fpubh.2022.874455.
- [26] P. K. Roy, A. Singh, "COVID-19 Disease Prediction Using Weighted Ensemble Transfer Learning," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 1, pp.13-22, 2023, doi:10.9781/ijimai.2023.02.006.
- [27] A. Andueza, M. Á. D. Arco-Osuna, B. Fornés, R. González-Crespo, J. M. Martín-Álvarez, "Using the Statistical Machine Learning Models

- ARIMA and SARIMA to Measure the Impact of Covid-19 on Official Provincial Sales of Cigarettes in Spain," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 1, pp. 73-87, 2023, doi:10.9781/ijimai.2023.02.010.
- [28] H. P. Cowley, M. S. Robinette, J. K. Matelsky *et al.* "Using machine learning on clinical data to identify unexpected patterns in groups of COVID-19 patients," *Scientific Reports*, vol. 13, no. 2236, 2023, doi:10.1038/s41598-022-26294-9
- [29] J.T. Hancock and T.M. Khoshgoftaar, "Survey on categorical data for neural networks," *Journal of Big Data*, vol. 7, no. 28, pp. 1-41, 2020, doi: 10.1186/s40537-020-00305-w.
- [30] W. S. A. Farizi, I. Hidayah, M. N. Rizal, "Isolation Forest Based Anomaly Detection: A Systematic Literature Review," *2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*, Semarang, Indonesia, 2021, pp. 118-122, doi: 10.1109/ICITACEE53184.2021.9617498.
- [31] N. Pudjihartono, T. Fadason, A.W. Kempa-Liehr and J.M. O'Sullivan, "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction," *Frontiers in Bioinformatics*, vol. 2, no. 927312, 2022, doi: 10.3389/fbinf.2022.927312.
- [32] K. Dissanayake and M. G. Md Johar, "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2021, no. 1, 2021, doi:10.1155/2021/5581806
- [33] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, H. Liu, "Feature Selection: A Data Perspective," *Association for Computing Machinery*, vol. 50, no. 6, pp. 1-45, 2017, doi: 10.1145/3136625.
- [34] V. V. Iyer and A. E. Yilmaz, "Using the ANOVA F-Statistic to Isolate Information-Revealing Near-Field Measurement Configurations for Embedded Systems," *2021 IEEE International Joint EMC/SI/PI and EMC Europe Symposium, Raleigh, NC, USA*, pp. 1024-1029, 2021, doi: 10.1109/EMC/SI/PI/EMCEurope52599.2021.9559360.
- [35] A. O. Odetunmbi, O. A. Adejumo, A. T. Anake, "A study of Hepatitis B virus infection using chi-square statistic," *Journal of Physics Conference Series*, vol. 1734, no. 01, 2021, doi:10.1088/1742-6596/1734/1/012010.
- [36] N. Carrara and J. Ernst, "On the estimation of mutual information," *Proceedings of The 39th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, vol. 33, no.1, 2020, doi:10.3390/proceedings2019033031.
- [37] I. M. Nasir, M. A. Khan, M. Yasmin, J. H. Shah, M. Gabryel, R. Scherer, R. Damaševičius, "Pearson Correlation-Based Feature Selection for Document Classification Using Balanced Training", *Sensors*, vol. 20, no. 23 pp. 1-18, 2020, doi : 10.3390/s20236793.
- [38] F. Saberi-Movahed, M. Mohammadifard, A. Mehrpooya, M. Rezaei-Ravari, K. Berahmand, M. Rostami, S. Karami, *et al.*, "Decoding Clinical Biomarker Space of COVID-19: Exploring Matrix Factorization-based Feature Selection Methods," *medRxiv [Preprint]*, 2021, doi: 10.1101/2021.07.07.21259699.
- [39] C. A. Ramezan, "Transferability of Recursive Feature Elimination(RFE)-Derived Feature Sets for Support Vector Machine Land Cover Classification," *Remote Sensing*, vol. 14, no. 24, 2022, doi: 10.3390/rs14246218.
- [40] C. Zhang, Y. Yi, L. Wang, X. Zhang, S. Chen, Z. Su, S. Zhang, Y. Xue, "Estimation of the Bio-Parameters of Winter Wheat by Combining Feature Selection with Machine Learning Using Multi-Temporal Unmanned Aerial Vehicle Multispectral Images," *Remote Sensing*, vol. 16, no. 3, pp. 1-22, 2024, doi:10.3390/rs16030469
- [41] A. Suruliandi, K. Ranjini, S. P. Raja, "Balancing Assisted Reproductive Technology Dataset for Improving the Efficiency of Incremental Classifiers and Feature Selection Techniques," *Journal of Circuits, Systems, and Computers, World Scientific*, vol. 30, no. 06, 2130007, 2021, doi:10.1142/S0218126621300075
- [42] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Computer Science*, vol. 2, no. 160, 2021, doi:10.1007/s42979-021-00592-x.
- [43] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189-215, 2020, doi: 10.1016/j.neucom.2019.10.118.
- [44] W. G. Gadallah, N. M. Omar and H. M. Ibrahim, "Machine Learning-based Distributed Denial of Service Attacks Detection Technique using New Features in Software-defined Networks," *International Journal of Computer Network and Information Security*, vol. 3, pp. 15-27, 2021, doi:10.5815/ijcnis.2021.03.02.
- [45] C. N. Villavicencio, J. J. E. Macrohon, X. A. Inbaraj, J. H. Jeng, J. G. Hsieh, "COVID-19 Prediction Applying Supervised Machine Learning Algorithms with Comparative Analysis Using WEKA," *Algorithms*, vol.14, no. 7, 2021, doi:10.3390/a14070201.
- [46] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR)[Internet]*, vol. 9, no. 1 pp.381-386, 2020.
- [47] M. Rohini, K. R. Naveena, G. Jothipriya, S. Kameshwaran, M. Jagadeeswari, "A Comparative Approach to Predict Corona Virus Using Machine Learning," *Proceedings of the International Conference on Artificial Intelligence and Smart Systems, Coimbatore, India*, 2021, pp. 331-337, doi: 10.1109/ICAIS50930.2021.9395827.
- [48] T. Rymarczyk, E. Kozłowski, G. Kłosowski, K. Niderla, "Logistic Regression for Machine Learning in Process Tomography," *Sensors*, vol. 19, no. 15, 2019, doi:10.3390/s19153400.
- [49] S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, S. Parasa, "On evaluation metrics for medical applications of artificial intelligence," *Scientific Reports*, vol. 12, no. 1. 2022, doi: 10.1038/s41598-022-09954-8.
- [50] R. A. Rayan, A. Suruliandi, S.P. Raja, H. B. F. David, "A survey on an analysis of big data open source datasets, techniques and tools for the prediction of corona virus disease," *Journal of Circuits, Systems and Computers*, vol. 32, no. 12, 2023, doi:10.1142/S0218126623300039.
- [51] J. White, S. D. Power, "k-Fold Cross-Validation Can Significantly Over-Estimate True Classification Accuracy in Common EEG-Based Passive BCI Experimental Designs: An Empirical Investigation," *Sensors (Basel)*, vol. 23, no. 13, 2023, doi: 10.3390/s23136077.



A. Suruliandi

A. Suruliandi is currently the Professor and Head at Manonmaniam Sundaranar University, Tamil Nadu, India. He previously worked as a lecturer at Kamaraj College, Tamil Nadu. With a Bachelor's degree in Electronics and Communication Engineering, a gold medal in Master's degree, and a Ph.D., he has published over 80 research papers and actively contributes to the academic community through lectures, presentations, and peer reviewing. His research focuses on nurturing young minds and guiding research scholars in exploring innovative ideas.



R. Ame Rayan

R. Ame Rayan is currently pursuing Ph.D. at Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India. Prior to pursuing her Ph.D., she worked as an Assistant Professor in the Computer Science Department at Holy Cross Home Science College in Thoothukudi, Tamil Nadu, India. She has completed her undergraduate studies at St. Mary's College in Thoothukudi, Tamil Nadu, India. She obtained her Master's degree in Computer Applications (MCA) from Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India.



S. P. Raja

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

Multiscale Attentional Squeeze-And-Excitation Network for Person Re-Identification

Tiancun Guo¹, Qiang Zhou^{1*}, Mingliang Gao¹, Gwanggil Jeon^{2*}, David Camacho³

¹ The Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000 (China)

² The Department of Embedded Systems Engineering, Incheon National University, Incheon, 22012 (South Korea)

³ The Department of Computer Systems Engineering, Technical University of Madrid, Madrid (Spain)

* Corresponding author: zhouqiang@sdu.edu.cn (Q. Zhou), ggjeon@gmail.com (G. Jeon)

Received 14 May 2024 | Accepted 29 October 2024 | Early Access 15 January 2025



ABSTRACT

In recent years, with the advancement of deep learning, person re-identification (Re-ID) has become increasingly significant. The existing person Re-ID methods primarily focus on optimizing network architecture to enhance Re-ID task performance. However, these methods often overlook the importance of valuable features in distinguishing Re-ID tasks, leading to reduced model efficacy in complex scenarios. As a solution, we utilize the attention mechanism to develop the lightweight multiscale Attentional Squeeze-and-Excitation Network (MASNet) that can distinguish between significant and non-significant features. Specifically, we utilize the SEAttention (SE) module to amplify important feature channels and suppress redundant ones. Additionally, the Spatial Group Enhance (SGE) module is introduced to enable networks to enhance semantic learning expression and suppress potential noise autonomously. We conduct comprehensive experiments on Market1501, MSMT17, and VeRi-776 datasets and cross-domain experiments on MSMT17 \rightarrow Market1501 to validate the model performance. Experimental results prove that the proposed MASNet achieves competitive performance across all experiments.

KEYWORDS

Attention Mechanism,
Cross-Domain,
Multiscale, Person Re-ID.

DOI: 10.9781/ijimai.2025.01.001

I. INTRODUCTION

PERSON re-identification (Re-ID) is to determine whether pedestrian images extracted from different cameras or different video clips taken from the same camera are the same person. In recent years, person Re-ID has become a pivotal element within intelligent surveillance systems and has received significant attention from the computer vision community. Previous works [1]–[4] have made significant progress in the person Re-ID task. Most approaches still utilize a backbone model initially designed for generic image classification tasks [5]. Recent works [6] illustrate that using different architectures leads to model performance differences. Yet, some works for neural architecture search are still designed based on the traditional neural architecture search (NAS) methods employed for general classification tasks [7], [8]. The traditional NAS is associated with high computational costs and lacks generality. Also, the non-compatibility between the search scheme and actual world training schemes results in suboptimal performance in person Re-ID.

Aiming at the above problems, the MSINet [9] employs a twin comparison mechanism to eliminate the class binding between the training and validation sets. This mechanism offers more suitable supervision for neural architecture search in person Re-ID. It achieves

compatibility between the search and real-world training schemes and improves the task's performance. Additionally, a multiscale interaction module is devised to facilitate mutual enhancement among multiscale features. Yet, person Re-ID is a complex and challenging task. The MSINet fails to adequately address scenarios where crucial feature channels have a more pronounced impact on the task. We draw inspiration from the multiscale interaction network (MSINet). Meanwhile, we propose incorporating an attention mechanism to guide the network in prioritizing the more influential feature channels. Concurrently, we empower the architecture to suppress insignificant feature channel information.

In this study, we construct the multiscale Attentional Squeeze-and-Excitation Network (MASNet) by incorporating the Squeeze-and-Excitation (SE) attention module and Spatial Group Enhance (SGE) module. We enhance the ability of the network to capture details in complex scenes, suppress background noise, and adjust features through attention mechanisms to focus on key features. Specifically, the contributions of this work can be summarized as follows:

- The MASNet is proposed to concentrate on the more essential feature information in person Re-ID. The MASNet can acquire more meaningful insights about pedestrians rather than being influenced by noise-disturbing features.

Please cite this article as:

T. Guo, Q. Zhou, M. Gao, G. Jeon, D. Camacho. Multiscale Attentional Squeeze-And-Excitation Network for Person Re-Identification, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 4, pp. 99-106, 2025, <http://dx.doi.org/10.9781/ijimai.2025.01.001>

- An SE module is adopted to learn each feature channel's significance autonomously. Meanwhile, an SGE module is introduced to generate an attention factor for each spatial position within each semantic group.
- With only 2.5M model parameters, extensive experiments conducted on several public datasets have verified that the proposed model surpasses existing methods in detection accuracy.

II. RELATED WORK

With the advancement of deep learning, person Re-ID tasks have garnered increased attention within the domain of computer vision [10]. Researchers have proposed numerous methodologies in the realm of person Re-ID to enhance performance. Among these, the attention mechanism has gradually emerged as a crucial element in Re-ID.

A. Person Re-Identification

The objective of person Re-ID is to ascertain whether images of a person depict the same individual. The same or different cameras can capture these images at different times. There has been widespread research on person Re-ID based on deep learning [11]–[15]. Methods in deep learning for the person Re-ID task generally fall into two categories: designing more efficient networks and acquiring additional prior knowledge. Luo *et al.* [16] suggested a baseline that relies solely on the global features of ResNet50 to fulfill task performance requirements.

Zhou *et al.* [4] introduced a lightweight omni-scale network (OSNet) to capture various spatial scales and encapsulate multiscale collaborative composite features. Likewise, Li *et al.* [17] explored an efficient network architecture through microarchitecture search. They introduced the Top-k Sample Search strategy to achieve a cost-effective search while avoiding potential local optimal results. Some approaches apply body structure and posture information for site detection or person normalization. For example, Li *et al.* [18] utilized Spatial Transformer Networks (STN) with spatial constraints to learn and locate a person with attitude changes. The FD-GAN [19] is proposed to utilize identity-related and posture-independent representations. The FD-GAN sidesteps the necessity for additional pose information and reduces computational costs. In addition, some approaches [20], [21] concentrate on improving network performance by optimizing the loss function to enhance its relationship with the instance. For example, Gu *et al.* [22] proposed AutoLoss-GMS to search for an improved loss function within the loss function space to aim for efficient and excellent person Re-ID. Chen *et al.* [23] designed a quadruplet loss function and proposed a quadruplet deep network. The network incorporates online hard negative mining to enhance the model's generalization ability. Alternative methods [24], [25] focus on designing part-based models. They aim to emphasize the prominence of the person. Sun *et al.* [26] employed the Multi-Head Self-Attention Module (MHSAM) to address background confusion and occlusion challenges. While performance has been enhanced, the computational burden remains considerable. In this work, we achieve competitive performance with a lightweight architecture at a lower computing cost.

B. Attention-Based Person Re-ID

Recent years have witnessed considerable success in attention mechanisms in computer vision [27]–[30]. Also, the attention mechanism plays an indispensable role in person Re-ID. A body part detector is utilized to acquire the characteristics of a person's body parts [31], [32]. The connectivity of key points is utilized to generate a mask for human body parts and emphasize the representation of the human body [31]. Nevertheless, these methods heavily depend on the accuracy of analytical models of the human body or pose estimators.

For video person Re-ID tasks, multiple methods [33], [34] investigated key time series frames using attention mechanisms. Additionally, there are methods to map 2D images into 3D spaces, facilitating pedestrian matching [35]. [36] introduces a point cloud matching (PCM) strategy to calculate the distance of multi-view convergence and allow for the differentiation of different individuals. Furthermore, Long Short-Term Memory (LSTM) is utilized to construct the motion dynamics of 3D tasks to simplify person matching [37]. A Reinforced Temporal Attention (RTA) based neural network architecture is proposed in [38]. It features a Long Short-Term Memory (CNN-LSTM) face-matching algorithm that utilizes an RGB-Depth conversion method. [39] employs the double attention mechanism to optimize and align features. This approach tackles the challenge of blurred vision in real-world scenarios. Chen *et al.* [40] proposed a network, named as ABD-Net. Spatial and channel attention are combined in the ABD-Net to directly learn a person's feature information from data and context. In SCSN [41], multiple attention models are cascaded to capture diverse cues. However, the complexity of cascading architectures poses a challenge in avoiding redundant information duplication, which leads to high computational costs. Our focus is on enhancing Re-ID's performance by implementing an attention strategy. Simultaneously, we achieve good performance without incurring undue computing costs.

III. METHODOLOGY

In this section, we delve into the details of the methods and modules utilized in the model. We first describe the work accomplished in the baseline (MSINet) [9] to facilitate comprehension. Following that, we elaborate on the details of the SEAttention (SE) module. Subsequently, we describe the Spatial Group Enhance (SGE) module. The structure of MASENet is shown in the Fig 1.

A. Baseline

1. Twins Contrastive Mechanism

The NAS is designed to adaptively search for the optimal network architecture for given data. In [9], defining the common model variable as α and the structure variable as β . In the search space σ , with the network layer i , β_i can manipulate the weighted value of individual operation o . The feature undergoes these operations iteratively. Ultimately, the final output is weighted and generated through the soft maximum of the operational output. Equation (1) describes the output.

$$f(x_i) = \sum_{o \in \sigma} \frac{\exp\{\beta_i^o\}}{\sum_{o' \in \sigma} \exp\{\beta_i^{o'}\}} \cdot o(x_i) \quad (1)$$

The model parameters are updated based on training results. Subsequently, the schema parameters are updated using validation results. Since the testing and validation datasets share identical categories, Re-ID requires distinct categories to be included in both the training and validation datasets. This discrepancy results in incompatibility between the search scheme and the actual training scheme, potentially leading to suboptimal results. MSINet incorporates the Twin Comparison mechanism (TCM). Two independent auxiliary memories v_{tr} and v_{ver} are employed to reserve training features and validation data. In each iteration, the training loss is initially computed using the training auxiliary memory to provide data for model updates. Given the feature f with the class tag a , the classification loss is expressed as Equation (2).

$$L_{tr}^{cl} = -\log \frac{\exp(f \cdot v_{tr}^a / \tau)}{\sum_{h=0}^{H_{tr}^n} \exp(f \cdot v_{tr}^h / \tau)} \quad (2)$$

where v_{tr}^a represents the memory features associated with the class a , and H_{tr}^n represents the sum number of classes in the training data.

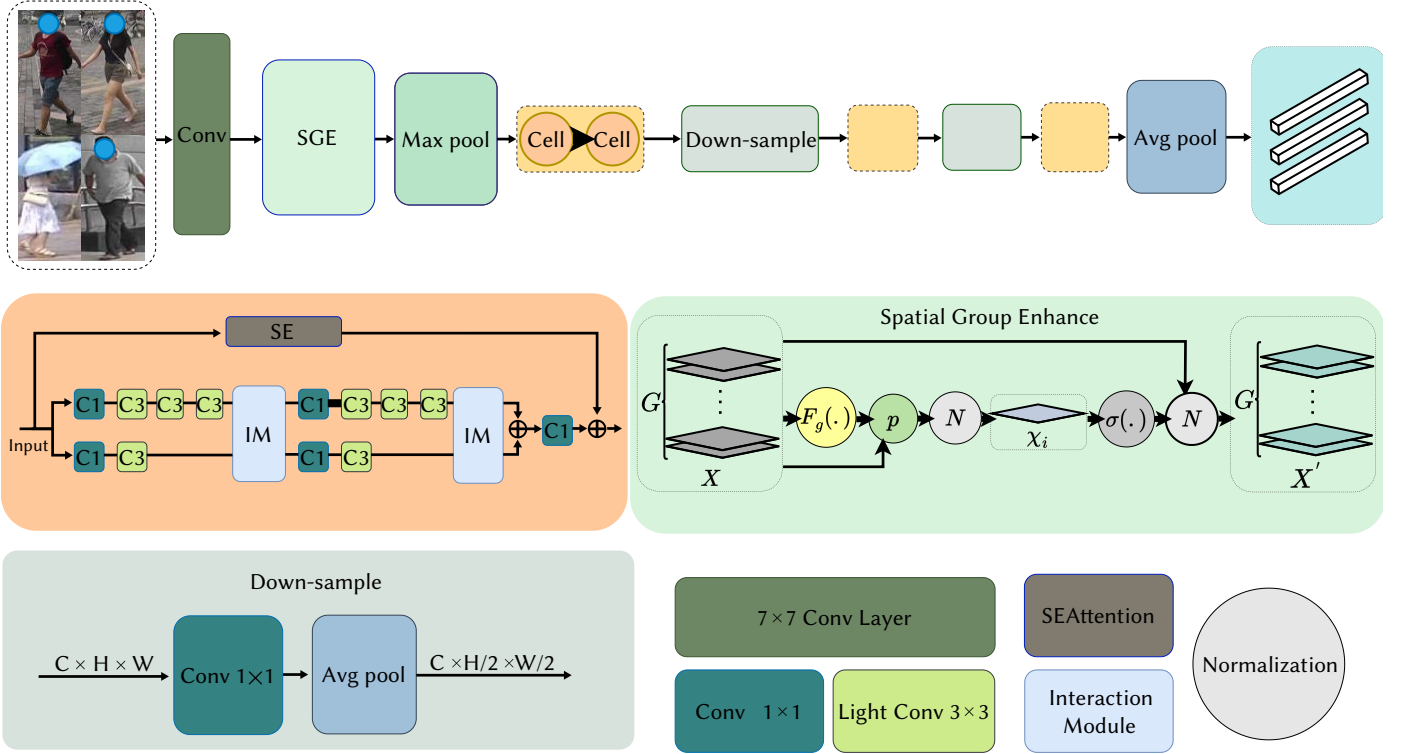


Fig. 1. The design of the proposed MASENet architecture. The MASENet allows for the input of pedestrian or vehicle images. The interaction module facilitates the exchange of information between two branches in each cell. The SE module enables the network to focus on useful feature channels, while the SGE enables each spatial group to enhance the expression of its learning autonomously.

The τ means the temperature argument should be set to 0.05 according to [42]. After the update, feature f is set to the corresponding memory feature by Equation (3).

$$v_{tr}^a \leftarrow \gamma v_{tr}^a + (1 - \gamma)f \quad (3)$$

where γ is set to 0.2 [42]. Substituting v_{ver} with v_{tr} produces validation losses and updates schema parameters. It completes the iteration by validating the loss update pattern parameters.

2. Multiscale Interaction Space

While previous Re-ID research has incorporated multiscale features, it was primarily designed based on experience. MSINet has devised a multiscale interaction space enabling features to interact with one another. As depicted in Fig. 1, features traverse two branches with different receptive field scales within each cell. To achieve a network with low computational complexity, a stack of multiple convolutions is employed to set the scale. The Interaction Module acts as a conduit for the interchange of features and information between the two branches. The IM can execute four operations on an input feature (y_1, y_2) : None. It does not involve operations with any parameters, yet accurately outputs (y_1, y_2) . Exchange. Considered one of the most powerful interactions, it can be directly interchanged between the two branches and (y_1, y_2) . Channel Gate. Channel Gate introduces Channel attention gates by Multi-Layer Perceptron (MLP) [43], [44], as shown in Equation (4):

$$G(y) = \delta(MLP(y)) \quad (4)$$

and returns value $(G(y_1) \cdot y_1, G(y_2) \cdot y_2)$. The MLP consists of two fully connected layers with parameters utilized by both branches. This enables networks to interact with each other by jointly filtering and validating feature channels.

Cross Attention Calculating the correlation between the two branches involves exchanging the keys of the branches. Then, the

correlated activation [45] is converted into a mask and appended to the original feature in the right proportion. As depicted in Fig. 1, the two branches are fused through summation operations after the interaction. It's crucial to highlight that the additional parameters introduced by multiple interaction modules are limited. Each unit can be searched in the context of the entire network without being impacted. The interactions that carry the most weight β_i^o for each layer are saved, thereby shaping the search architecture. After the architecture search, the model undergoes training to incorporate classification ID loss and triple loss, as shown in Equation (5):

$$L_{ID} = \frac{1}{N} \sum_{i=1}^N -\log\left(\frac{\exp M_i^T f_i}{\sum_i \exp M_i^T f_i}\right) \quad (5)$$

where f_i is features array, M_i is the relevant classifier weight. The triple loss is expressed as Equation (6).

$$L_{TRI} = [D(f_a, f_b) - D(f_a, f_n) + \rho]_+ \quad (6)$$

where f_a, f_b, f_n are the inlaid features of the anchor. $D(f_a, f_b), D(f_a, f_n)$ represent the Euclidean distance. ρ is the edge argument. $[\cdot]_+$ means the $\max(\cdot, 0)$ function.

B. Structure of SEAttention Module

Learning extensive feature information solely through convolution kernels and achieving high performance is quite challenging for person Re-ID. Hence, we introduce the SEAttention module. From the perspective of feature channel information, SE specifies channel interdependencies without significantly increasing the network's depth or width. This technique results in only an increase in the number of model parameters. SE does not significantly increase the network's computational complexity. The importance weight of each feature channel can be adjusted based on its varying importance to the network. The network autonomously learns importance weights to enhance crucial feature channels and suppress redundant ones.

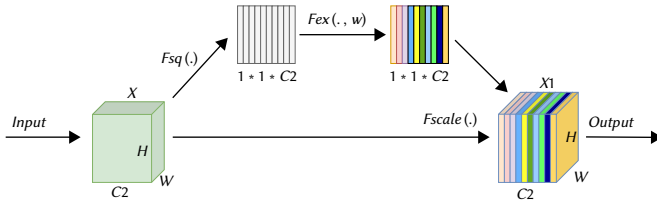


Fig. 2. The model structure of the Squeeze-and-Excitation (SE).

As depicted in Fig. 2, the SE module is integrated into residuals to accentuate the more significant feature channels. The SE module operation is divided into three steps: First, acquire each feature channel's global compression feature through global average pooling. Secondly, the new weight value of each feature channel will be derived from 0 to 1 via two fully connected layers. Lastly, matrix multiplication of the new weight value with the original feature channel will be performed using the SE module's feature channel recalibration function. Then, the output of the two branches is weighted and combined with the output of the SE module after a 1×1 conv.

C. Structure of Spatial Group Enhance Module

Feature representations of objects are generated by convolutional neural networks (CNNs) by acquiring semantic sub-features at different levels. Yet, the activation of these sub-features is often influenced by spatial noise. Therefore, we introduce the SGE [46] module to generate attention factors for each spatial position in each semantic group, shown in Fig. 1. It helps the module adjust each sub-feature's importance and suppress potential noise. Specifically, SGE divides feature graphs into groups G along channel dimensions. Each individual group has vectors representing each position in space, as shown in Equation (7):

$$X = \{x_1 \dots x_M\}, x_i \in \mathbb{R}^{\frac{C}{G}}, M = H \times W \quad (7)$$

where C is the number of channels. Within this group space, the network can learn the feature representation of the key region. Unlike CNN, which struggles to obtain uniformly distributed features, SGE utilizes global statistical features through the spatial average function to approximate the semantic vectors learned by the group. The Equation (8) is as follows:

$$F_g(X) = \frac{1}{M} \sum_{i=1}^M x_i \quad (8)$$

Then, the global features are utilized to generate an importance weight value for each feature. This weight value is obtained by (9):

$$p_i = F_g(X) \cdot x_i \quad (9)$$

It is worth noting that Eq. (9) can be reformulated as shown in Equation (10):

$$p_i = |F_g(X)| |x_i| \cos(\delta_i) \quad (10)$$

where δ_i is the angle between $F_g(X)$ and x_i . We apply spatial normalization to p to avoid bias amplitude discrepancies between samples [47], [48]. It is mathematically expressed Equation (11).

$$p'_i = \frac{p_i - \eta_c}{\phi_c + \varepsilon}, \eta_c = \frac{1}{M} \sum_j p_j, \phi_c^2 = \frac{1}{M} \sum_j (p_j - \eta_c)^2 \quad (11)$$

where ε is a constant added for numerical stability. To ensure that normalization in the network can also represent the identity exchange, a pair of parameters φ, λ is introduced into each coefficient p'_i . The formula for scaling and moving normalized values is shown in Equation (12):

$$\chi_i = p'_i \varphi + \lambda \quad (12)$$

where the quantity of what φ, λ is the same as the number of G . Finally, the original x_i is scaled by the generated importance coefficients χ_i through a sigmoid function gate $\sigma(\cdot)$ over the space, as shown in Equation (13):

$$x'_i = x_i \cdot \sigma(\chi_i) \quad (13)$$

Then, the enhanced feature vectors will be obtained, and the element group will be formed with these enhanced feature vectors. The specific form is given by Equation (14).

$$X' = \{x'_1 \dots x'_M\}, x_i \in \mathbb{R}^{\frac{C}{G}}, M = H \times W \quad (14)$$

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

The MASENet is tested on two Re-ID datasets about pedestrians: Market1501 [49], MSMT17 [50]. To assess the model's generalization ability, the MASENet is also evaluated on VeRi-776 [51], [52] and MSMT17 \rightarrow Market1501 [49]. For simplicity and convenience, the three datasets are named M, MS, and VR. The output evaluation indexes are common performance metrics for person Re-ID, including mean average precision (mAP) and cumulative matching features (CMC).

B. Comparative Experiments With Other Lightweight Network

We initially contrast MASENet with the recently proposed lightweight network by in-domain and cross-domain experiments. The results in the table are pre-trained on ImageNet.

In-Domain Test. The initial learning rate is set at 0.065. During training, the learning rate is adjusted at epochs 150, 225, and 300. We use a Stochastic Gradient Descent (SGD) optimizer with a momentum coefficient of 0.9 and a weight decay of 0.0005. The parameters are updated using triple loss and cross-entropy loss. The value of p in formula (6) is set to 0.3. Adopting the same structure as CDNet [17], and the specific experimental results are shown in Table I.

ResNet50 is the most common backbone network for person Re-ID, but it performs the worst on the three datasets mentioned in this

TABLE I. THE PERFORMANCE ON RE-ID DATASETS. THE RESULTS ARE PRE-TRAINED ON IMAGENET IN ADVANCE

Method	Params	M		MS		VR		MS \rightarrow M	
		Rank-1 \uparrow	mAP \uparrow	Rank-1 \uparrow	mAP \uparrow	Rank-1 \uparrow	mAP \uparrow	Rank-1 \uparrow	mAP \uparrow
ResNet50* [16]	~24M	94.5	85.9	75.5	50.4	94.5	73.6	58.8	31.8
OSNet [44]	2.2M	94.8	84.9	78.7	52.9	95.5	76.4	66.6	37.5
CDNet [17]	1.8M	95.1	86.0	78.9	54.7	-	-	-	-
MSINet [9]	2.3M	95.3	89.6	81.0	59.6	96.8	78.8	74.9	46.2
MSINet-SAM [9]	2.4M	95.5	89.9	80.7	59.5	96.7	79.0	76.3	48.4
MASENet (Ours)	~2.5M	95.9	89.9	81.9	60.8	95.9	79.5	77.3	50.1

* represents the results reproduced by the baseline

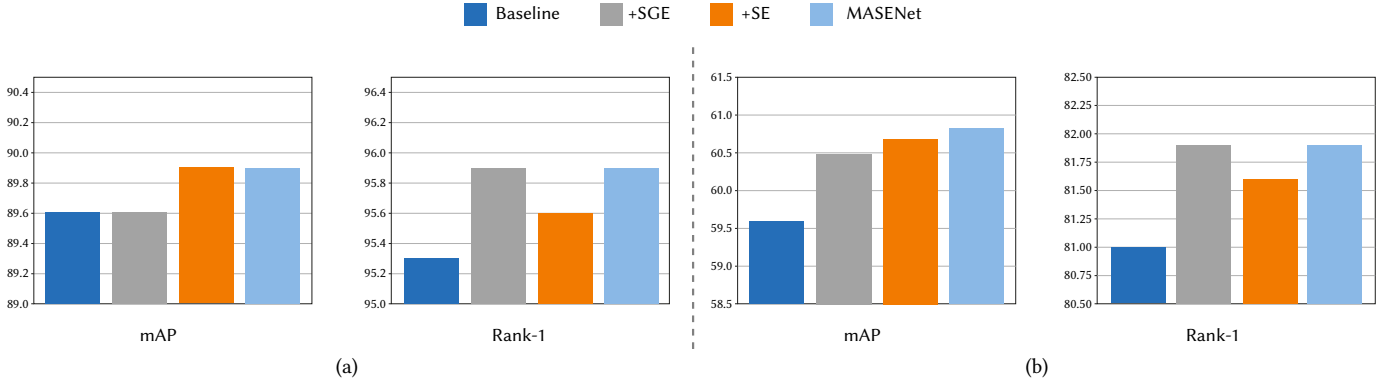


Fig. 3. Ablation experiments on (a) Market1501 dataset; (b) MSMT17 dataset.

article. Additionally, ResNet50 heavily relies on ImageNet pre-training operations. Unlike other datasets, the MS dataset presents more complex situations, such as background noise and attitude changes. The style of the MS dataset is more in line with real-world application scenarios. To overcome the challenges of complex scenarios, MASENet integrates SE and SGE modules. The SE module adjusts the importance of each feature channel adaptively to allow the network to learn and emphasize key features. The SGE module produces attention factors for each spatial location to adjust subfeature importance and mitigate background noise. These introduced modules enhance the network's feature representation ability and capture the detailed elements of complex scenes more effectively. Compared to MSINet, the MASENet improved mAP and Rank-1 by 1.3% and 1.2%. The results on MS validate that MASENet is more effective at handling complex scenarios and focusing on more important feature channels than the baseline. OSNet [44] and CDNet [17] are recent architectures designed for Re-ID, both addressing the issue of multiscale feature fusion. CDNet utilizes the traditional NAS scheme for searching. Table II shows the optimal interaction within each cell. It shows that the MASENet outperforms most lightweight networks.

TABLE II. THE DETAIL ABOUT INTERACTION OPERATION. N: NONE; E: EXCHANGE; G: CHANNEL GATE; C: CROSS ATTENTION

Cell.1		Cell.2		Cell.3		Cell.4		Cell.5		Cell.6	
1	2	3	4	5	6	7	8	9	10	11	12
G	G	E	G	C	G	G	N	G	C	E	C

The model employed for person gender Re-ID is applied to the VR dataset for experiments. Table I indicates that in the VR experiment, mAP has increased by 1.2%. This case signifies an enhancement in the model's processing capability for generally complex scenes.

Cross-Domain Test. Cross-domain experiments are commonly employed to assess the generalization ability of models. MASENet is pre-trained with 250 epochs and fine-tuned to prevent overfitting. Table 1 demonstrates that ResNet50 is susceptible to image styles. The efficient interaction of MSINet can be effectively extended to other image domains. To enhance the generalization ability of MSINet, [9] introduced the spatial alignment module (SAM) module to align spatial correlation between person images. Yet, the performance of the proposed network on $MS \rightarrow M$ shows a substantial improvement compared to the MSINet-SAM. The results that mAP and Rank-1 are respectively up 1.7% and 1% than MSINet-SAM, further demonstrating the significant enhancement the module brings to the model.

C. Comparative Results With State-of-Art Methods

Table III offers additional insight into the supervised performance contrast between the proposed MASENet and SOTA methods on M and MS datasets. MASENet successfully achieves the objective of high

TABLE III. THE PERFORMANCE CONTRAST BETWEEN MASENET AND SOTA METHODS ON MARKET1501 AND MSMT17 DATASETS

Method	M		MS	
	Rank-1 \uparrow	mAP \uparrow	Rank-1 \uparrow	mAP \uparrow
Auto-ReID+ [55]	95.8	88.2	80.8	59.5
RGA-SC [54]	96.1	88.4	80.3	57.5
BAT-Net [56]	95.1	87.4	79.5	56.8
SFT [57]	94.1	87.5	79.0	58.3
CARL [53]	95.8	89.2	-	-
DRL-Net [58]	94.7	86.9	78.4	55.3
GCN [59]	95.3	85.7	-	-
PAT [60]	95.4	88.0	-	-
C2F [61]	94.8	87.7	-	-
BoT [62]	94.5	85.9	-	-
MGN* [63]	95.7	86.9	76.9	52.1
ISP [64]	95.3	88.6	-	-
OSNet [44]	93.6	81.0	71.0	43.3
CDNet [17]	95.1	86.0	78.9	54.7
MSINet [9]	95.3	89.6	81.0	59.6
MASENet (Ours)	95.9	89.9	81.9	60.8

precision with reduced computational requirements. The proposed method achieved an mAP of 89.9% and a Rank-1 accuracy of 95.9% on the Market1501 dataset.

Similarly, on the MS dataset, the proposed method achieves an mAP of 60.8% and a Rank-1 accuracy of 81.9%. CARL [53] introduces a measure of camera pairing loss for learning. Compared with CARL, the proposed method improves mAP and Rank-1 by 0.1% and 0.7% on the M dataset. It is worth noting that compared with MS dataset, M dataset has a simple style and certain limitations. During the training process of MASENet, the advantages brought by further feature enhancement may be difficult to fully exert. This situation may result in the limited performance improvement of the proposed method in the M dataset. Additionally, RGA-SC [54] incorporates a relation-aware global attention module. On the MS dataset, MASENet outperforms with a 1.6% boost in Rank-1 accuracy and a 4% enhancement in mAP. Although MASENet's Rank-1 performance on the M dataset is slightly lower than that of the RGA-SC method, it still demonstrates near-optimal performance. It verifies the effectiveness of matching top-ranked predictions.

Evaluation of the challenging MSMT17 dataset reveals that the proposed network also possesses the ability to handle challenging scenarios.

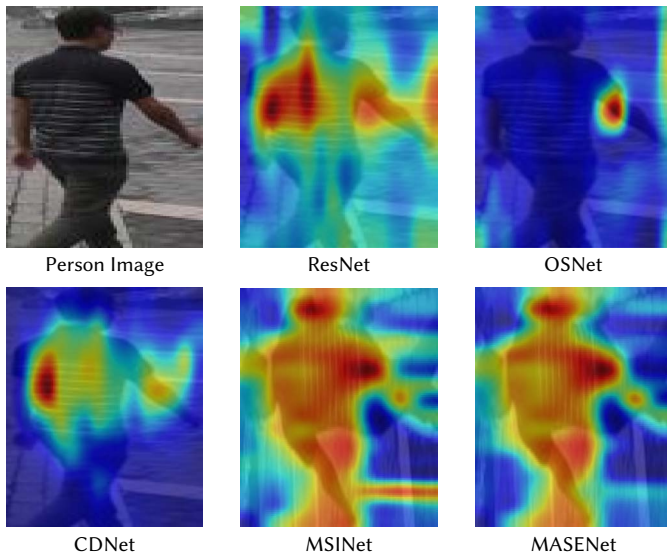


Fig. 4. Attention maps are generated from baseline (MSINet), OSNet, CDNet, and the proposed MASENet.

The attention maps are shown in Fig. 4. It proves that the introduced modules aid in filtering out inconsequential background noise, thereby enhancing the model's focus on the critical features of the pedestrian.

D. Ablation Studies

The performance improvement of MASENet primarily stems from the inclusion of SE and SGE modules. In this section, we conduct ablation experiments to validate the effectiveness of each module in enhancing network performance. The detailed results can be found in Table III. Additionally, we visualize the output from the baseline with independently introduced SE and SGE modules. As illustrated in Fig 5, the incorporation of SE and SGE modules effectively accentuates personal features while suppressing background noise interference. **Baseline.** Compared with other methods, baseline conducts neural architecture searches through twin comparison mechanisms. An effective interactive module also enables information exchange between two branches. These results, from the Market1501 dataset, with a map of 89.6% and Rank-1 with 95.3%, and from the MSMT17 dataset, with a map of 59.6% and Rank-1 with 81.0%, illustrate the improvement of model performance. However, real-world situations are intricate, and background noise can affect the model's performance in person Re-ID tasks. The performance of the baseline on MS suggests that it has not experienced significant improvement compared to other methods.

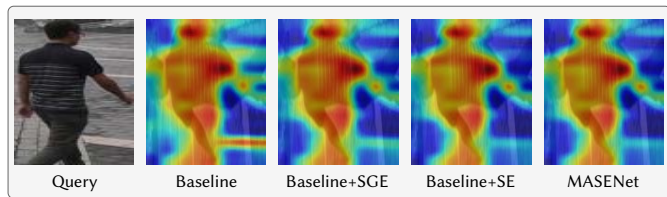


Fig. 5. Visualization of network output. The same sample is selected as in Fig 4. The baseline refers to MSINet.

SE Module. SE autonomously learns the importance of each feature channel in the feature map and assigns a weight value to enhance important feature channels. Fig 3 illustrates that SE improves the network's performance on both M and MS datasets. In particular, MASENet improves mAP and Rank-1 by 1.1% and 0.6% on the MS dataset. This demonstrates that the SE's attention to important feature channels effectively enhances the model's accuracy for person retrieval.

SGE Module. SGE generates attention factors for each spatial position in each semantic group. This capability empowers the network to independently enhance the expression of spatial semantic learning and suppress potential noise. Notably, the feature space enhancement mechanism proves especially advantageous for CMC. However, improvements in mAP are influenced by multiple factors, including dataset characteristics and optimization strategies. Moreover, the enhancement mechanism of the SGE module might alter feature distribution, potentially introducing deviations that impact the mAP performance. Overall, the SGE module is considerably more effective than the SE module in improving the network's cumulative matching feature.

V. CONCLUSION

In this work, we proposed a baseline approach for architectural search and incorporated the attention mechanism to create MASENet. Specifically, we introduce the SEAttention module to improve the network's attention to valuable feature channels. The Spatial Group Enhance module is introduced to enhance the expression of spatial semantic learning and suppress noise. This equips the network to address person Re-ID tasks with more complex backgrounds and poses. Experimental results demonstrate that MASENet exhibits outstanding performance and generalization ability on both person Re-ID and vehicle datasets. In the future, further optimization based on the CNN network architecture and SGE module will be explored. Additionally, the application of lightweight architecture and the enhancement of generalization performance will be pursued to adapt to complex Re-ID tasks.

ACKNOWLEDGMENT

This work has been funded by Grants: PLEC2021-007681 (XAI-DisInfodemics), PID2020-117263GB-I00 (FightDIS), and PCI2022-134990-2 (MARTINI) of the CHISTERA IV Cofund 2021 program, funded by MCIN/AEI/10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR"; by Calouste Gulbenkian Foundation, under the project MuseAI - Detecting and matching suspicious claims with AI, and by "Convenio Plurianual with the Universidad Polit'cnica de Madrid in the actuation line of Programa de Excelencia para el Profesorado Universitario".

REFERENCES

- [1] Y. Dai, J. Liu, Y. Bai, Z. Tong, L.-Y. Duan, "Dual- refinement: Joint label and feature refinement for unsupervised domain adaptive person re-identification," *IEEE Transactions on Image Processing*, vol. 30, pp. 7815–7829, 2021.
- [2] B. Yang, J. Chen, M. Ye, "Towards grand unified representation learning for unsupervised visible-infrared person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11069–11079.
- [3] C. Zhao, X. Lv, Z. Zhang, W. Zuo, J. Wu, D. Miao, "Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3180–3195, 2020.
- [4] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, "Learning generalisable omni-scale representations for person re-identification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5056–5069, 2021.
- [5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
- [6] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE transactions on*

- pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 2872–2893, 2021.
- [7] H. Liu, K. Simonyan, Y. Yang, “Darts: Differentiable architecture search,” *arXiv preprint arXiv:1806.09055*, 2018. Available online: <https://arxiv.org/abs/1806.09055>.
 - [8] X. Dong, Y. Yang, “Searching for a robust neural architecture in four gpu hours,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1761–1770.
 - [9] J. Gu, K. Wang, H. Luo, C. Chen, W. Jiang, Y. Fang, S. Zhang, Y. You, J. Zhao, “Msinet: Twins contrastive search of multi-scale interaction for object reid,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19243–19253.
 - [10] G. Zou, G. Fu, X. Peng, Y. Liu, M. Gao, Z. Liu, “Person re-identification based on metric learning: a survey,” *multimedia tools and applications*, vol. 80, no. 17, pp. 26855–26888, 2021.
 - [11] S. Liao, S. Z. Li, “Efficient psd constrained asymmetric metric learning for person re-identification,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3685–3693.
 - [12] Y. Liu, G. Zou, G. Chen, M. Gao, L. Yin, “Unsupervised person re-identification based on distribution regularization constrained asymmetric metric learning,” *Applied Intelligence*, vol. 53, no. 23, pp. 28879–28894, 2023.
 - [13] B. Chen, W. Deng, J. Hu, “Mixed high-order attention network for person re-identification,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 371–381.
 - [14] H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan, “End-to-end comparative attention networks for person re-identification,” *IEEE transactions on image processing*, vol. 26, no. 7, pp. 3492–3506, 2017.
 - [15] G. Chen, T. Gu, J. Lu, J.-A. Bao, J. Zhou, “Person re-identification via attention pyramid,” *IEEE Transactions on Image Processing*, vol. 30, pp. 7663–7676, 2021.
 - [16] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, J. Gu, “A strong baseline and batch normalization neck for deep person re-identification,” *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2597–2609, 2019.
 - [17] H. Li, G. Wu, W.-S. Zheng, “Combined depth space based architecture search for person re-identification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6729–6738.
 - [18] D. Li, X. Chen, Z. Zhang, K. Huang, “Learning deep context-aware features over body and latent parts for person re-identification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 384–393.
 - [19] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, *et al.*, “Fd-gan: Pose-guided feature distilling gan for robust person re-identification,” *Advances in neural information processing systems*, vol. 31, 2018.
 - [20] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, “Person re-identification by multi-channel parts-based cnn with improved triplet loss function,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1335–1344.
 - [21] Y. Yuan, J. Zhang, Q. Wang, “Deep gabor convolution network for person re-identification,” *Neurocomputing*, vol. 378, pp. 387–398, 2020.
 - [22] H. Gu, J. Li, G. Fu, C. Wong, X. Chen, J. Zhu, “Autoloss- gms: Searching generalized margin-based softmax loss function for person re-identification,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4744–4753.
 - [23] W. Chen, X. Chen, J. Zhang, K. Huang, “Beyond triplet loss: a deep quadruplet network for person re-identification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 403–412.
 - [24] G. Chen, J. Lu, M. Yang, J. Zhou, “Spatial-temporal attention-aware learning for video-based person re-identification,” *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4192–4205, 2019.
 - [25] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, J. Sun, “Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 393–402.
 - [26] H. Tan, X. Liu, B. Yin, X. Li, “MHSA-Net: Multihead self-attention network for occluded person re-identification,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8210–8224, 2023.
 - [27] X. Guo, M. Gao, W. Zhai, J. Shang, Q. Li, “Spatial-frequency attention network for crowd counting,” *Big data*, vol. 10, no. 5, pp. 453–465, 2022.
 - [28] S. Jiao, J. Wang, G. Hu, Z. Pan, L. Du, J. Zhang, “Joint attention mechanism for person re-identification,” *IEEE Access*, vol. 7, pp. 90497–90506, 2019.
 - [29] R. Quan, X. Dong, Y. Wu, L. Zhu, Y. Yang, “Auto- reid: Searching for a part-aware convnet for person re-identification,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3750–3759.
 - [30] W. Zhai, M. Gao, Q. Li, G. Jeon, M. Anisetti, “Fpanet: feature pyramid attention network for crowd counting,” *Applied Intelligence*, vol. 53, no. 16, pp. 19199–19216, 2023.
 - [31] J. Xu, R. Zhao, F. Zhu, H. Wang, W. Ouyang, “Attention-aware compositional network for person re-identification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2119–2128.
 - [32] C. Song, Y. Huang, W. Ouyang, L. Wang, “Mask- guided contrastive attention model for person re-identification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1179–1188.
 - [33] G. Chen, J. Lu, M. Yang, J. Zhou, “Learning recurrent 3d attention for video-based person re-identification,” *IEEE Transactions on Image Processing*, vol. 29, pp. 6963–6976, 2020.
 - [34] G. Chen, Y. Rao, J. Lu, J. Zhou, “Temporal coherence or temporal motion: Which is more critical for video- based person re-identification?,” in *Computer Vision– ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII* 16, 2020, pp. 660–676, Springer.
 - [35] Z. Zheng, X. Wang, N. Zheng, Y. Yang, “Parameter-efficient person re-identification in the 3d space,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 6, pp. 7534–7547, 2024.
 - [36] M. Munaro, A. Fossati, A. Basso, E. Menegatti, L. Van Gool, “One-shot person re-identification with a consumer depth camera,” *Person Re-Identification*, pp. 161–181, 2014.
 - [37] A. Haque, A. Alahi, L. Fei-Fei, “Recurrent attention models for depth-based person identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1229–1238.
 - [38] N. Karianakis, Z. Liu, Y. Chen, S. Soatto, “Person depth reid: Robust person re-identification with commodity depth sensors,” *arXiv preprint arXiv:1705.09882*, 2017.
 - [39] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, G. Wang, “Dual attention matching network for context-aware feature sequence based person re-identification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5363–5372.
 - [40] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, Z. Wang, “Abd-net: Attentive but diverse person re-identification,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8351–8361.
 - [41] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, Y. Yang, “Saliency-guided cascaded suppression network for person re-identification,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3300–3310.
 - [42] Y. Ge, F. Zhu, D. Chen, R. Zhao, *et al.*, “Self-paced contrastive learning with hybrid memory for domain adaptive object re-id,” *Advances in neural information processing systems*, vol. 33, pp. 11309–11321, 2020.
 - [43] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, “Cbam: Convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
 - [44] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, “Omni- scale feature learning for person re-identification,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3702–3712.
 - [45] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, “Dual attention network for scene segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.
 - [46] X. Li, X. Hu, J. Yang, “Spatial group-wise enhance: Improving semantic feature learning in convolutional networks,” *arXiv preprint arXiv:1905.09646*, 2019. Available online: <https://arxiv.org/abs/1905.09646>.
 - [47] Y. Wu, K. He, “Group normalization,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
 - [48] S. Qiao, H. Wang, C. Liu, W. Shen, A. Yuille, “Micro-batch training with batch-channel normalization and weight standardization,” *arXiv preprint arXiv:1903.10520*, 2019. Available online: <https://arxiv.org/abs/1903.10520>.

- [49] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.
- [50] L. Wei, S. Zhang, W. Gao, Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 79–88.
- [51] X. Liu, W. Liu, T. Mei, H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, 2016, pp. 869–884.
- [52] X. Liu, W. Liu, H. Ma, H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *2016 IEEE international conference on multimedia and expo (ICME)*, 2016, pp. 1–6.
- [53] J. Wu, Y. Yang, Z. Lei, Y. Yang, S. Chen, S. Z. Li, "Camera-aware representation learning for person re-identification," *Neurocomputing*, vol. 518, pp. 155–164, 2023.
- [54] Z. Zhang, C. Lan, W. Zeng, X. Jin, Z. Chen, "Relation-aware global attention for person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3186–3195.
- [55] H. Gu, G. Fu, J. Li, J. Zhu, "Auto-reid+: Searching for a multi-branch convnet for person re-identification," *Neurocomputing*, vol. 435, pp. 53–66, 2021.
- [56] P. Fang, J. Zhou, S. K. Roy, L. Petersson, M. Harandi, "Bilinear attention networks for person retrieval," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8030–8039.
- [57] C. Luo, Y. Chen, N. Wang, Z. Zhang, "Spectral feature transformation for person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4976–4985.
- [58] M. Jia, X. Cheng, S. Lu, J. Zhang, "Learning disentangled representation implicitly via transformer for occluded person re-identification," *IEEE Transactions on Multimedia*, vol. 25, pp. 1294–1305, 2022.
- [59] G. Xie, X. Wen, L. Yuan, H. Xu, Z. Liu, "Global correlative network for person re-identification," *Neurocomputing*, vol. 469, pp. 298–309, 2022.
- [60] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2898–2907.
- [61] A. Zhang, Y. Gao, Y. Niu, W. Liu, Y. Zhou, "Coarse- to-fine person re-identification with auxiliary-domain classification and second-order information bottleneck," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 598–607.
- [62] H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 1487–1495.
- [63] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 274–282.
- [64] K. Zhu, H. Guo, Z. Liu, M. Tang, J. Wang, "Identity- guided human semantic parsing for person re-identification," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 2020, pp. 346–363, Springer.



Tiancun Guo

Tiancun Guo is pursuing an M.S. degree at the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His research interests include person re-identification, deep learning, and computer vision.



Qiang Zhou

Qiang Zhou (corresponding author) received his Ph.D. degree from Chongqing University, Chongqing, China, in 2016. He is an associate professor in the Department of Smart Grid Information Engineering, School of Electrical and Electronic Engineering, Shan Dong University of Technology, Shandong, China. His current research interests include artificial intelligence, image recognition, and smart power grid.



Mingliang Gao

Mingliang Gao received his Ph.D degree in Communication and Information Systems from Sichuan University. He is now an associate professor and vice dean at the Shandong University of Technology. He was a visiting lecturer at the University of British Columbia during 2018–2019. His research interests include computer vision, machine learning, and intelligent optimal control. He has published over 150 journal/conference papers in IEEE, Springer, Elsevier, and Wiley.



Gwanggil Jeon

Gwanggil Jeon (corresponding author) received a Ph.D. degree from the Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea, in 2008. From 2009.09 to 2011.08, he was with the School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada, as a Post-Doctoral Fellow. Dr. Jeon is an IEEE Senior Member, a recipient of the IEEE Chester Sall Award in 2007, the ETRI Journal Paper Award in 2008, and the Industry-Academic Merit Award by the Ministry of SMEs and Startups of Korea Minister in 2020. He serves as a full professor at Shandong University of Technology, Zibo, China, and Incheon National University, Incheon, Korea. His research interests include computer vision, machine learning, and the Internet of Things.



David Camacho

David Camacho is full professor at Computer Systems Engineering Department of Universidad Politécnica de Madrid (UPM), and the head of the Applied Intelligence and Data Analysis research group (AIDA: <https://aida.etsisi.uam.es>) at UPM. He holds a Ph.D. in Computer Science from Universidad Carlos III de Madrid in 2001 with honors (best thesis award in Computer Science). He has published more than 300 journals, books, and conference papers. His research interests include Machine Learning (Clustering/Deep Learning), Computational Intelligence (Evolutionary Computation, Swarm Intelligence), Social Network Analysis, Fake News and Disinformation Analysis. He has participated/led more than 50 research projects (Spanish and European: H2020, DG Justice, ISFP, and Erasmus+), related to the design and application of artificial intelligence methods for data mining and optimization for problems emerging in industrial scenarios, aeronautics, aerospace engineering, cybercrime/cyber intelligence, social networks applications, or video games among others. He has served as Editor in Chief of Wiley's Expert Systems since 2023 and sits on the Editorial Board of several journals, including Information Fusion, IEEE Transactions on Emerging Topics in Computational Intelligence (IEEE TETCI), Human-centric Computing and Information Sciences (HCIS), and Cognitive Computation among others.

Automatic Surveillance of People and Objects on Railway Tracks

Domingo Martínez Núñez¹, Fernando Carlos López Hernández², J. Javier Rainer Granados³ *

¹ Central Control Station, Metro de Madrid 28029, Madrid (Spain)

² Applied Mathematics Department, Universidad Complutense de Madrid (UCM), 28040, Madrid (Spain)

³ Universidad Internacional de La Rioja (UNIR) 26006, Logroño (Spain)

* **Corresponding author:** domingo.martinez@metromadrid.es (D. Martínez Núñez), felh@ucm.es (F. C. López Hernández), javier.rainer@unir.net (J. J. Rainer Granados).

Received 30 October 2023 | Accepted 19 July 2024 | Early Access 20 August 2024



ABSTRACT

This paper describes the development and evaluation of a surveillance system for the detection of people and objects on railroad tracks in real time. Firstly, the paper evaluates several background subtraction techniques including CNNs and the object detection library called YOLO. Then we describe a novel strategy to mitigate the occlusion caused by the perspective of the camera and the integration of an alarms and pre-alarms policy. To evaluate its performance, we have implemented and automated the control and notification aspects of the surveillance system using computer vision techniques. This setup, running on a standard PC, achieves an average frame rate of 15 FPS and a latency of 0.54 seconds per frame, meeting real-time expectations in terms of both false alarms and precision in operational mode. The results from experiments conducted with a publicly available recorded video dataset from Metro de Madrid facilities demonstrate significant improvements over current state-of-the-art solutions. These improvements include better accident anticipation and enhanced information provided to the operator using a standard low-cost camera. Consequently, we conclude that the approach described in this paper is both effective and a more practical, cost-efficient alternative to the other solutions reviewed.

KEYWORDS

Computer Vision,
Machine Learning,
Neural Networks,
Railway Safety,
Surveillance.

DOI: 10.9781/ijimai.2024.08.004

I. INTRODUCTION

CURRENTLY modern railroad facilities use video cameras that transmit signals to a limited set of monitors, usually called CCTV (Closed Circuit Television), for the detection of people and objects in dangerous situations. They include manual surveillance and alarms that can interrupt the circulation of trains and warn the security services. For this kind of automatic surveillance problems, well-known computer vision (CV) techniques, including object detection and image background subtraction, as well as neural network classification are frequently implemented in other problem domains.

This research contributes to improving railway safety. In addition to the integration of current computer vision techniques and deep learning algorithms, we describe a system capable of detecting in real time the presence of people and objects on the tracks, in a way that overcomes the traditional methods' limitations. The relevance of this work lies in its ability to offer an economical and effective alarm system, which relies on the use of a single low-cost camera per monitored area, thus boosting its technical and economic convenience over other more costly or complex approaches.

Our research contributes to railroad security, as it combines two innovative techniques to achieve a more practical and economically efficient surveillance, using only a low-cost single camera per monitored area (480x640), and using an alarm policy effectively.

This system contributes to the minimization of accidents that disturb the smooth running of railroad lines, due to people or objects falling on the tracks, as well as violations of railroad rules during service hours (suicides, graffiti, vandalism, crossing of tracks, assaults, etc.).

The automatic processing of the camera images lets us monitor in a more continuous and systematic way than humans could without interruptions, avoiding visual fatigue [1], distinctions, and thus letting the personnel to attend to other duties. Our system is semi-automatic because it reports potentially dangerous situations alerting the security station staff or traffic controllers, who will be able to stop the trains.

The use of affordable equipment is practical and beneficial for successful railroad surveillance. While researchers have reported successful approaches using expensive equipment such as LiDAR, laser, or multiple cameras for object detection on railroads [2], these

Please cite this article as:

D. Martínez Núñez, F. C. López Hernández, J. J. Rainer Granados. Automatic Surveillance of People and Objects on Railway Tracks, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 4, pp. 107-116, 2025, <http://dx.doi.org/10.9781/ijimai.2024.08.004>

methods are often cost-prohibitive. Few studies have focused on achieving affordable railroad object detection using only a low-cost single camera with CV techniques. Existing affordable single-camera CV techniques [3][4] are primarily designed for tracking rails rather than detecting objects and people for video surveillance purposes on railroads. Our research aims to address this gap by demonstrating how a low-cost single camera can be effectively used for real-time detection of both objects and people on the tracks.

We address solutions for challenges inherent to railway surveillance, such as the variability of lighting conditions and the complexity of the scenarios on the tracks. In addition, the challenge of efficiently processing the large volume of data generated by cameras is significant. Our approach provides solutions for integrating highly efficient image processing algorithms capable of discriminating between false alarms and real risk situations. This integration not only improves the accuracy in detecting dangerous situations, but also optimizes the use of surveillance resources, resulting in a safer and more efficient operation of railway facilities.

Additionally, this study conducts a review of current approaches in existing rail surveillance installations, identifying their limitations in terms of operational costs and effectiveness. Recognizing these limitations, we have designed a system that improves the detection and classification of objects and people and effectively integrates into existing rail surveillance operational requirements. Our methodology integrates existing deep learning technology with computer vision techniques, into the constraints and operational needs of a real rail facility. The proposed solution maintains a balance between efficiency and affordability, making it a viable and attractive solution for a wide range of railway applications.

II. STATE OF THE ART

In the literature there are different proposed techniques to detect people or objects on the railroad track platforms, being relatively expensive and difficult to maintain technology (in comparison to our solution). S. Oh et al. [5] propose using multiple cameras perpendicular along the track to monitor almost the entire length of the track line of the platform. The author divides the line monitoring process into two parts: detection of train status and detection of objects and people on the track.

Other approaches [6][7] that incorporate different sensing modalities, such as LiDAR, laser or remote sensing data, have been developed to monitor pathways.

T. Xiao et al. [8] have developed a non-contact multisensory technology detection technique. In particular, they have deployed an On-Board obstacle detection device based on a camera and a LiDAR to detect obstacles in the monitored area in real-time, and to determine whether the train should brake automatically, or the train pilot should brake manually.

S. Taori et al. [9] have gone one step farther by suggesting the implementation of a multisensory barrier composed of infrared (IR) and ultrasonic (US) sensors, in addition to a CV system, in order to alert the surveillance system about the presence of obstacles on the train track and thus prevent possible accidents.

In recent years, various not as expensive methods have been developed to track the rails. Note that this problem does not coincide exactly with our surveillance problem of detecting falling objects and people on the tracks. For example, F. Kaleli and Y. S. Akgul [10] proposed an algorithm based on dynamic programming to track the trajectory of the front part of the train on the railroad. This method consists of three steps: first, a Sobel operator is used to identify the borders of the input image, then a Hough transform is applied

to the binary image to detect the railroad line, and finally dynamic programming is used to efficiently track the rail lines.

Y. Wang et al. [11] developed a neural network approach called RailNet, which includes a segmentation network to track the rails, trained with their own railroad segmentation dataset, that is, a collection of annotated images specifically designed for identifying and delineating various elements of railroad infrastructure. This approach also includes a feature extraction network and a segmentation network.

M. Ghorbanalivakili et al. [12] proposed a rail path extraction process in which the pixels of the left-right rails of each path are extracted and associated using a convolutional architecture called TPE-Net. This net has two different regression branches to get the locations of the center points of each rail and generate the possible train routes (called "ego-routes"). The experimental results indicate that this technique has a high accuracy and recovery of true positives.

Recently authors such as M. Qasim Gandapur and E. Verdú [13], M. Adimoolam et al. [14] and A. D. Petrović et al. [15] have proposed a combination of Convolutional Neural Network (CNN)s based on YOLO-v5 [16] implementation, for object detection and tracking with CV. The later also includes a Canny edge detection and the Hough transform.

Another technique named Mask RCNN [17] employs a pyramidal structure to obtain high performance in the task of instance segmentation on the COCO dataset [18].

This last approach is similar to ours, but it differs in that our research does not focus on detecting traffic signals and track bifurcation, but on detecting falling objects and people on the tracks.

H. Pan et al. [19] propose a multi-task learning network that segment, detects, and classifies the rail lines. This approach makes improvements over other multitask networks such as paying more attention to accuracy than to recall. Segmentation aids in improving the classification results. The railroad detection algorithm in the multitask network can effectively deal with the blocked track line problem, and it avoids the disadvantage of segmentation network in detecting tracking lines with thinner and smaller pixel ratio. Its anchorless designs avoid the problem that exists when the size of the anchor frame is not suitable for small targets.

Recently at London Underground's Docklands Light Railway (DLR), experimentation is underway with the development of the CCTV AI Trial project [20]. The alert system has been deployed to prevent accidents on the network lines and strives to minimize false alarms. This pilot project is in the testing phase.

In the field of low-light image processing in railway driving, the paper presented by Z. Chen et al. [21] describes an innovative network designed to optimize visual clarity in low-light conditions. They describe a progressive enhancement strategy and a lightweight network overcoming the limitations of conventional methods, delivering remarkable results that outperform previous techniques such as Zero-DCE++, SCI, RetinexDIP, and RUAS. With its efficient structure combining advanced feature extraction operators and accurate encoder-decoder architecture, these authors enhance image analysis for railway applications, contributing to the safe and reliable operation of trains in low-light conditions.

The work of C. Meng et al. [22] proposes SDRC-YOLO as an enhancement of the YOLOv5s algorithm, specifically designed to identify infiltrations in railway scenarios. It incorporates a Hybrid Attention SSA mechanism that combines a Spatial Attention Module (SAM) with Squeeze and Excitation Network (SENet) channel attention. The structure features a DW-Decoupled Head for efficiency and employs Large Convolutional Kernels from RepLKNet with strong

parameterization to create wider perceptual fields. Additionally, the lightweight universal resampling operator CARAFE is used to select more suitable sizes and proportions for infiltration features.

Recent advances in railway track segmentation algorithms have shown remarkable improvements in intrusion detection and overall security. A prominent proposal is ERTNet, an efficient railway track region segmentation algorithm based on a lightweight neural network and cross-fusion decode as described by Z. Chen et al. [23]. This network incorporates encoder-decoder architecture, using depth convolutions and cross-fusion to efficiently integrate shallow and deep features, achieving high accuracy with a lightweight model. They archive a MIoU (Mean Intersection over Union) of 92.4% with minimal computational resource requirements. ERTNet represents an advance in railway surveillance technology, and aligns with our system objectives to improve real-time and ensure the integrity of railway infrastructure.

Another significant development in railway region segmentation is the LRseg [24] model, designed to optimize the detection of foreign objects on tracks. This model incorporates a lightweight coding approach and a self-correcting decoder together with a segmentation head, enabling efficient, real-time processing, crucial for applications in on-board devices. With its low parameter requirements and its ability to operate efficiently on both embedded systems and powerful personal computers, LRseg represents a new contribution to railway image segmentation, delivering accurate and fast results, indispensable for railway safety and maintenance.

Another important advance in object detection on railway lines is the development of the RailFOD23 [25] dataset, specifically designed to improve automated detection of foreign objects such as plastic bags, flying objects, bird nests and balloons. This dataset includes 14,615 detailed annotated images generated using artificial intelligence techniques. Therefore, it represents a valuable resource for training object detection models, with direct applications in railway safety. This dataset promises to facilitate significant advances in railway surveillance technology, optimizing detection and response to potential threats in the power transmission infrastructure.

Despite advancements in railway surveillance, current systems exhibit significant limitations that justify the development of our proposed system. First, most systems rely on multiple cameras or expensive technologies like LiDAR, increasing complexity and implementation costs. Second, many of these systems are not optimized to operate in real-time on standard computing equipment, limiting their applicability in conventional railway environments. Additionally, the accuracy in detecting objects and people on tracks is often limited, leading to a high number of false positives and negatives. These limitations highlight the need for a more efficient and economical system, such as the one we propose, which uses a single low-cost camera and advanced algorithms for effective and real-time surveillance.

III. MATERIALS AND METHODS

Our real-time railway surveillance system employs a two-step detection process involving background subtraction and object classification using a convolutional neural network (CNN), specifically designed to detect people and objects on the tracks. By means of background subtraction it identifies changes in the scene to flag potential intruders or left objects, while the CNN classifies the detected objects to minimize false positives.

The MOG2 algorithm is used for background detection due to its capacity to extract dynamic backgrounds and its ability to recover the background state efficiently, which is crucial for the fast-moving

environment of a railway. It distinguishes between the background and the moving objects effectively, even for small or slow-moving items. MOG2 has a high processing speed, thanks to the CUDA¹ implementation.

The YOLO-v5 neural network model was selected for object classification because of its high accuracy and speed, vital for real-time processing. It identifies and classifies objects as either trains or people on the tracks, contributing to the alarm system's accuracy.

The system is implemented on standard PC hardware, achieving an average frame rate of 15 FPS with a latency of 0.54 seconds per frame, ensuring it meets real-time operation criteria. The alarm system distinguishes between actual threats and non-threats, issuing alerts for the security personnel to act accordingly.

Our system is divided into two detection steps, the background subtraction and the classification of objects in images through a CNN, integrated to form a cohesive system. This methodological approach is crucial to understanding the innovations and results of the study.

To evaluate which background subtraction technique best suits the needs of the system, several tests and iterations have been performed with the latest techniques. In particular, we have compared in terms of accuracy and image processing speed several algorithms, namely: MOG2 [26] (Gaussian Mixture Based Background/First Plane Segmentation Algorithm), GMG [27] (Combination of statistical estimation of the background image and Bayesian segmentation per pixel), KNN [28] (Nearest Neighbors), and CNT [29] (Count Based). Regarding image classification and based on the study of the existing literature and our preliminary tests of different neural networks that perform this task, YOLO-v5 has been chosen. YOLO-v5 has also been compared with its previous versions regarding accuracy and speed for image processing.

Regarding the evaluation of the background subtraction algorithm, we have measured its accuracy, i.e., how many errors the algorithm makes when identifying the background in an image or video. This has been done by comparing the results of the algorithm with a ground truth, i.e., reference images/videos that have already been manually labeled as the correct background.

Another aspect to evaluate the background subtraction algorithm is to measure its processing speed, i.e., how long it takes to process an image or video frame. This is important because real-time performance is required.

We conducted a series of tests to validate the efficiency of the system. These tests included varied scenarios, from low-light conditions to the presence of moving objects at different speeds. We continuously adjusted the algorithm's parameters to optimize its performance, paying particular attention to minimizing false positives and improving real-time detection.

This iterative process of testing and fine-tuning was key to adjusting the system to the specific conditions of the train tracks, ensuring effective and reliable surveillance.

To ensure a thorough understanding of our algorithm, below we provide a more detailed description of its components. Initially, the background subtraction stage employs the MOG2 algorithm, selected for its ability to effectively adapt to dynamic changes in the environment. This is a fundamental feature in the fast-moving railway context. This process facilitates the initial identification of potentially dangerous objects and people on the tracks. Subsequently, object classification is performed using the pre-trained YOLO-v5 model, which allows for a clear distinction between different types of objects

¹ CUDA (Compute Unified Device Architecture) Provides a development environment for creating high-performance GPU-accelerated applications. It includes GPU-accelerated libraries, debugging and optimization tools, a C/C++ compiler and a runtime library.

and checks that alerts are only issued for true dangerous situations. This workflow is illustrated in Fig. 1, which provides a simplified visual representation of the process from image capture to alert generation.

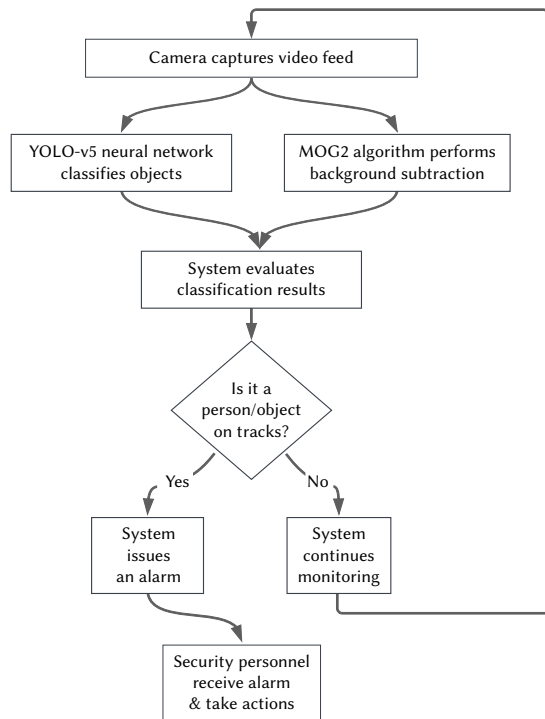


Fig. 1. Functional Diagram.

For the YOLO-v5 algorithm there exists comparisons and evaluations with different metrics already published². As can be seen in the AUC-ROC³ curve plot in Fig. 2, the YOLOv5x6 model with the YOLOv5x6.pt weights previously trained from pytorch.org/hub/, is the one that shows the best performance in speed and mAP⁴ [30] on the COCO [18] dataset of 5000 images at various inference sizes from 256 to 1536.

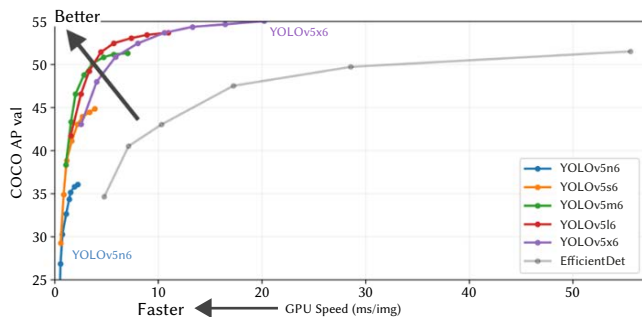


Fig. 2. Metrics of different weights in the COCO val2017 dataset.

During implementation, we faced significant technical challenges. One of the main challenges was the variability of environmental conditions affecting detection accuracy. To overcome this, we implemented adaptive algorithms that adjusted detection parameters based on lighting and weather conditions.

² Source code available at <https://github.com/ultralytics/yolov5>

³ AUC-ROC (Area Under the Curve - Receiver Operating Characteristic) is a metric used to evaluate the ability of a model to discriminate between positive and negative classes.

⁴ mAP: Mean precision (mAP) is the average of all classes of the maximum precision for each object at each recall value.

Another challenge was differentiating between irrelevant objects and real threats. To address this, we enhanced the algorithm's classification capability, teaching it to recognize a wider range of objects and situations. This improvement resulted in a significant reduction in false positives.

These adjustments and enhancements were crucial in ensuring the system's effectiveness and reliability in a real operational environment.

For the evaluation of the classification, the selection criteria for the videos took into account the limited availability of material, as these videos are from the actual operation of Metro de Madrid S.A. (the company in charge of Madrid's underground railway). We prioritized videos with a variety of environmental conditions, such as different lighting levels, to test the robustness of the system. In addition, videos were selected that included various types of objects and people on the tracks, ensuring that they did not reflect serious accident situations and various configurations of the camera installation, height, angle, etc. This selection allowed us to evaluate the effectiveness of the system in a wide range of realistic scenarios.

To ensure the reproducibility of our study for other researchers, we have made available the source code, scripts and the video dataset used in GitHub⁵ and Zenodo⁶ repositories respectively.

In this study, special emphasis was placed on the selection and preparation of the data set, which is crucial for the evaluation of our system.

The dataset used in our research comes from various video hosting platforms. This approach ensures a varied and realistic representation of the situations that could be encountered on railroad tracks. During the preprocessing and cleaning process, we focused on removing redundant elements, such as repetitions and labels, to optimize the data for analysis. This includes the removal of audio, thus improving the quality of the dataset. In addition, we have taken rigorous measures to comply with copyright regulations, ensuring respect for intellectual property and proper citation of all sources used.

As for the labeling of the data, a manual classification of the videos was carried out, from those presenting accidents of greater to lesser simplicity, highlighting the effort and meticulousness in the labeling process. However, we faced several challenges and limitations with this dataset. Critical aspects such as inherent biases, imbalance in class distribution, the presence of noise in the data, and the scarcity of videos related to the specific subject matter of the study were discussed. In addition, special emphasis was placed on ethical considerations, such as privacy, consent to data use, and anonymization, to ensure a responsible and ethical approach to our research.

A. Metrics Used for Background Subtraction

To evaluate background subtraction, the number of non-zero pixels per frame has been taken as a metric, because it indicates the effectiveness of the algorithm in detecting moving objects. A high number of non-zero pixels indicates that the algorithm has been able to effectively detect moving objects and separate them from the static background. On the other hand, a low number of pixels different from zero, indicates that the algorithm has had difficulty separating the moving objects from the static background.

To evaluate the background subtraction algorithm, it is necessary to measure its processing speed, since processing time efficiency is vital, especially in applications that require fast response, such as in railway surveillance systems. Different algorithms can have significant variations in their execution times, which directly affects the practicality of their application in real-time environments.

⁵ https://github.com/Domy5/Rail_Surveillance.

⁶ <https://doi.org/10.5281/zenodo.8357129>

Therefore, a balance between detection accuracy and processing speed is essential in choosing the most suitable algorithm for our system.

Our system leverages the processing capabilities of GPUs, where available, to significantly accelerate real-time video analysis. By relying on YOLOv5, the object detection model, we are guaranteeing current state of the art efficiency and speed. This model, together with the MOG2 background subtraction technique, allows the system to operate efficiently and with low energy consumption even on standard PC hardware. In addition, the choice of inexpensive cameras already installed for data collection not only makes our solution cost-effective, but it also contributes to the reduction of energy consumption, a key factor in continuous surveillance systems.

B. Metrics Used for Object Classification

In the evaluation of our object classifier model, we have employed the confusion matrix, a standard tool in the analysis of classification models. This matrix allows us a detailed understanding of the effectiveness of the model, breaking down the results into TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative). To provide a more complete view of model performance, we have supplemented these metrics with precision, sensitivity (or true positive rate) and specificity (or true negative rate). These additional measures help us evaluate the model's ability to correctly identify threats while minimizing false alarms, a crucial aspect of effective rail surveillance systems.

C. Global Metrics

The metric we used to evaluate the computational performance of the complete system (both the object classifier and the background subtraction technique) is the processing time latency of each image, which determines if it is appropriate for use in a real-time system.

We also analyzed the overall effectiveness and viability of the system through direct tests with the dataset videos. In particular, we reviewed the overall system performance and verified the result of the alarms and pre-alarms complied with the expected result.

IV. ALARM POLICY

In this section, we evaluate the system while it is operating on a regular basis, without incidents that affect the tracks, or that are not triggered by false negatives. These tests were performed to ensure that the system does not generate false alarms or unnecessary pre-alarms. This is an important measure to ensure the reliability of the system and to avoid saturation or desensitization of operators to real alarms.

Our system's alarm policy has been evaluated to maximize accuracy in identifying real risk situations while minimizing disruptions caused by false positives. We have developed an innovative approach that distinguishes between alarms and pre-alarms based on the severity and probability of the detected threat. Alarms are triggered when a person or object on the tracks is identified with high certainty, while pre-alarms are issued in situations of lower certainty, allowing for additional verification prior to raising the alarm. This strategy ensures a rapid and effective response to real threats and significantly reduces unnecessary operational disruptions. Below we describe the alarm policy and the logical implementation decisions for activating or deactivating alarms.

A. Alarm per Person on Track Platform

One of the challenges of our solution, using the existing CCTV cameras, is to determine if a person is on the platform or on the road, that is, if a person is within what corresponds to the ROI.

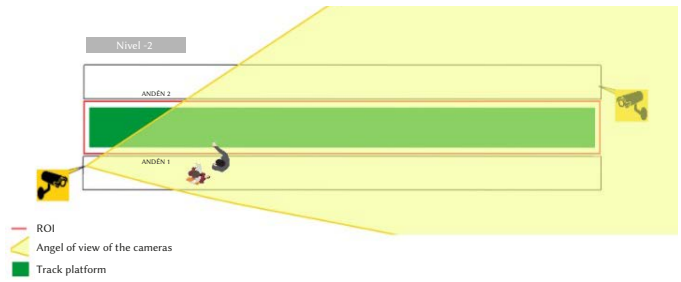


Fig. 3. Perspective of the camera with respect to the track platform.

As shown in Fig. 3, the location of the cameras is at the end of each platform, with an angle that focuses on an oblique angle to the end of the opposite platform. For this reason, the cameras generate images with a perspective that causes the occlusion of a certain part of the track platform by people near the edge of the platform where the camera is located.

If a person is located at the edge of the platform, the object detection system will generate a BBox or bounding box that will cross the area of interest or ROI as shown in the example in Fig. 4, being a false alarm, as the person is not on the track platform, but on the platform.



Fig. 4. BBox generated by object detection system.

To prevent the occlusion caused by people at the edge of the platform, with respect to the ROI, we decided to generate a point as far away as possible, this being at the bottom right of each BBox. This point corresponds to the green point in left foot of the person facing the camera in the example in Fig. 5. This point will serve as our reference to activate the alarm per object inside the ROI, and so emitting an alert sound. This means that when the reference point is inside the polygon that represents the ROI, the person on the track alarm will be triggered.



Fig. 5. Point generated at the bottom right of each BBox (left foot of each person).

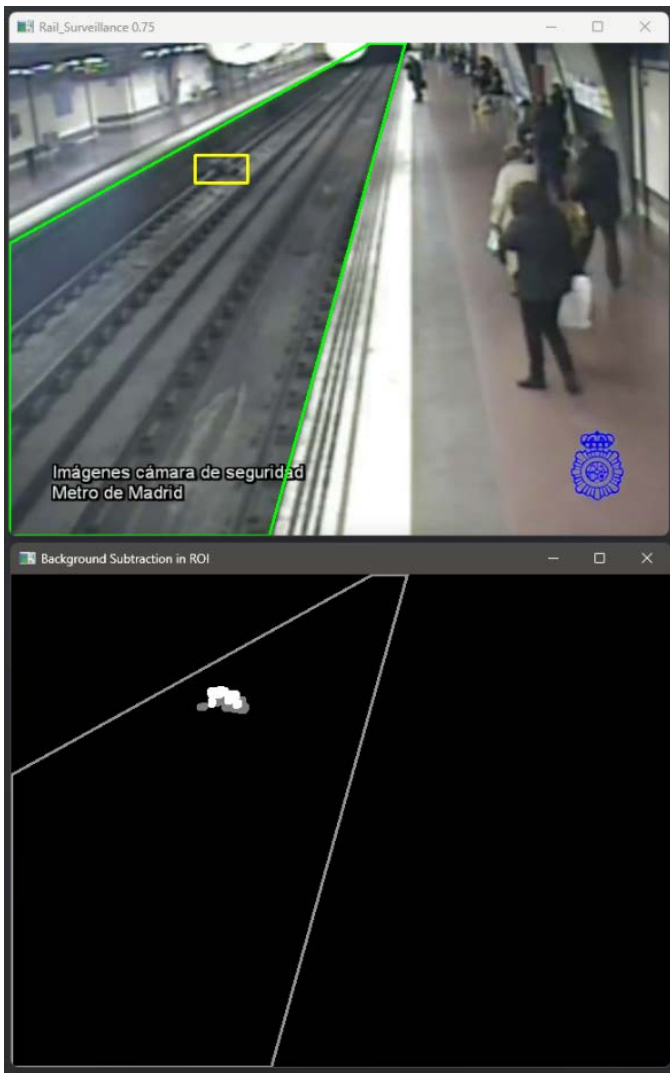


Fig. 6. Detection by background subtraction within the ROI.

B. Pre-alarms Due to Detecting Movements on the Railroad

The MOG2 background subtraction algorithm generates pre-alarms when a consecutive series of frames has a significant number of non-zero pixels in the contour mask in Fig. 6. The pre-alarm emits an alert sound different and softer than the alarms. Background subtraction is performed only within the ROI of the mask in Fig. 6, which means that the movements on the platform are ignored. This method detects events such as falling people, bulky objects, flash flooding of the track basin, vault detachment, etc.

C. Deactivating Alarms on Train Arrival

Once the arrival of the train is detected, the pre-alarms are deactivated, but not the alarms generated by person on the track platform (i.e., people within ROI). This type of alarm, in addition to emitting the alert sound, will display the operator the message: "Arrival of the train with person on the train track", as shown in Fig. 7.

V. RESULTS

In this section we describe and analyze the results of the two parts of the system separately: background subtraction and object classification. Subsequently the whole system is evaluated.

To evaluate the energy efficiency and fast processing of our system, we have performed benchmark tests using a standard PC configuration

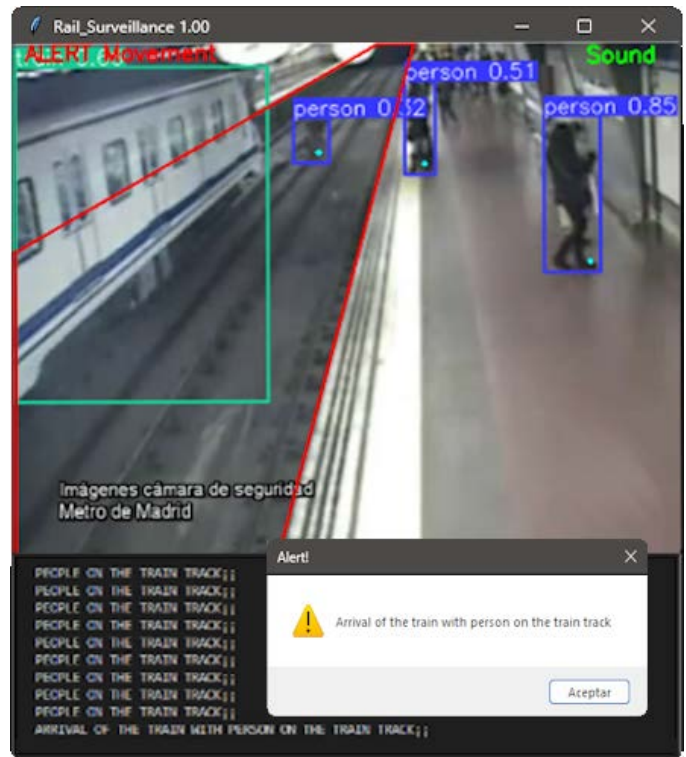


Fig. 7. Person on track message.

without additional specialized hardware. The results showed that the system maintains an average of 15 FPS and a latency per frame of 0.54 seconds, meeting real-time operation expectations and excelling in processing efficiency. This performance is due to the optimized system design and strategic use of GPU technology to accelerate computer vision tasks.

A. Evaluation of the Background Subtraction Algorithms

To evaluate the background subtraction algorithms, we compared the accuracy of the CNT, KNN, GMG and MOG2 algorithms. One of the main differences between them is that MOG2 uses an adaptive Gaussian blending model, while GMG uses a global Gaussian blending model. This means that MOG2 can adapt dynamically to changes in the scene background, allowing for greater accuracy in background identification and better performance in situations with a dynamic background. In comparison, the GMG algorithm uses the first frames of a video to build a model of the background, making it less adaptable to changes in the scene. Another important difference is that MOG2 is able to handle foreground object overlaps, so it is able to identify slowly moving objects in the scene. We have found that the GMG algorithm produces more artifacts and noise than MOG2. Therefore, there are significant differences between them that make MOG2 better than GMG to detect people moving slowly in dangerous situations.

CNT is based simply on pixel count, KNN only takes into account the distance between pixels and may work less accurately in situations with a dynamic background. MOG2, compared to KNN, is more flexible, as it allows easier adjustment of its parameters. Therefore, MOG2 achieves higher accuracy in identifying small objects.

We have tested several configurations in the parameters that let us modify the behavior of each subtraction algorithm and, in those subtraction algorithms that permit it, the application of the morphological aperture kernel to find the best result with respect to noise.

Fig. 8 shows the number of non-zero pixels per frame, which indicates that the algorithm has been able to effectively detect moving objects and separate it from the background. The train arrival occurs

at frame 740, each element or line shows the values of the number of pixels different from zero detected by each subtraction algorithm. As can be seen in the figure, the results are very similar in terms of number of pixels with the CNT and KNN algorithms (dark blue and yellow line, respectively). In both algorithms it takes many frames to recover the background state, which can lead to false positives events. GMG (gray line) has strong constant fluctuations that generate a lot of noise in the mask. The higher stability and speed in image changes can be observed in the MOG and MOG2 algorithms (light blue and orange line).

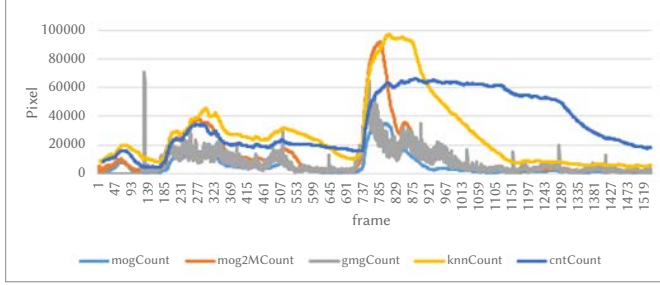


Fig. 8. Number of non-zero pixels per frame.

Fig. 9 shows the frame processing time of each algorithm. The best performance is achieved by the CNT algorithm with 0.0010 seconds average time, followed by MOG2, with 0.0032 seconds average time, being MOG2 more stable than KNN (0.005 seconds), GMG (0.015 seconds), and MOG (0.007 seconds), whose execution times double the value of MOG2.

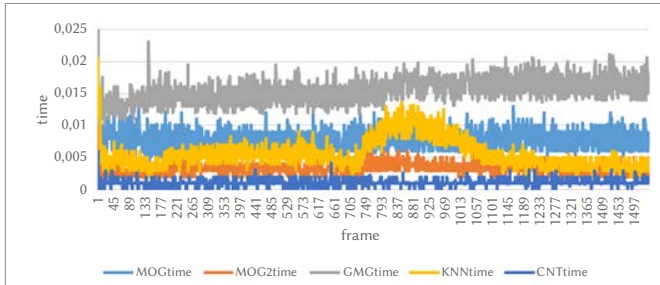


Fig. 9. Processing time per frame.

B. Object Classifier

To evaluate the object classifier, several of the videos available in the test dataset were chosen. The frames were labeled manually, then we executed the classifier to get the results for the “Train object” and “Person ontrack” event used to generate the confusion matrix in Table I and Table II.

TABLE I. CONFUSION MATRIX OF THE “TRAIN OBJECT” EVENT

Object: TRAIN		PREDICTION	
		POSITIVES	NEGATIVES
REAL	POSITIVES	792	4 (Type II error)
	NEGATIVES	8 (Type I error)	732

Next, several metrics are extracted from the confusion matrix for the “Train object” event in Table I. In particular, the precision is 99%, the accuracy is 99.21%, the specificity is 98.91%, the Recall or sensitivity is 99.49%, and the F1 score is 99.24%.

Overall, these values indicate that the classifier model performs well in classifying the “Train object” classification dataset.

TABLE II. CONFUSION MATRIX OF THE OBJECT “PERSON ON TRACK” EVENT

OBJECT: PERSON ON VIA		PREDICTION	
		POSITIVES	NEGATIVES
REAL	POSITIVES	29	504 (Type II error)
	NEGATIVES	0 (Type I error)	1003

Table II shows the confusion matrix for the “Person on track” dataset. In particular, the precision is 100%, the accuracy is 67.18%, the specificity is 100%, the Recall or sensitivity is 0.05%, and the F1 score is 0.10%.

These results indicate that the model has limited performance in classifying objects of the “Person on track” class. Usually this happens because the model tends to be over-fitted to the “Person on track” class, resulting in high accuracy, but limited generalization performance.

C. Overall Evaluation of the System

The first evaluation reported in Fig. 10 shows the latency inference per frame using eGPU with the test dataset.

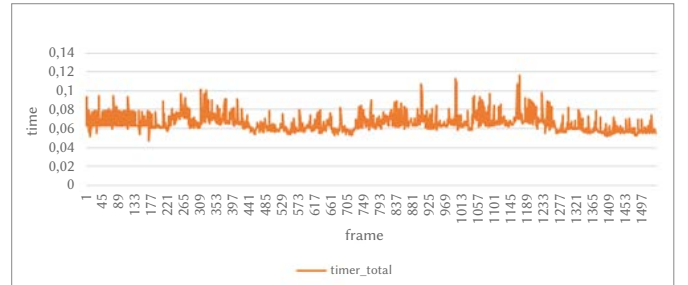


Fig. 10. Per-frame inference latency with eGPU.

The Fig. 11 shows the FPS data achieved by the system.

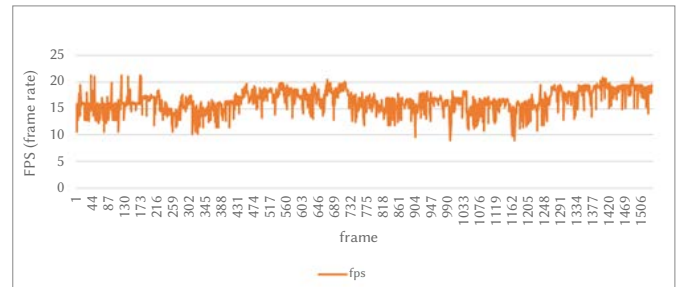


Fig. 11. FPS achieved with eGPU.

The results of the total frame time latency evaluations by inference indicate that the video surveillance system operates correctly for most of the video fragments analyzed, maintaining acceptable performance when processing incoming data with latency not exceeding 100 milliseconds. This confirms that the system meets the established standards to operate efficiently within the time limits required for real-time applications.

Finally, the system was evaluated in several video fragments in normal exploitation that is without events. In these fragments, false alarms do not occur. This means that the system is working properly and meets the operational expectations.

Our system has been rigorously evaluated to ensure its effectiveness and efficiency in a real operating environment. The accuracy of our background subtraction algorithm was validated by comparing the number of non-zero pixels per frame, revealing accurate detection of moving objects against the static background. This analysis

demonstrated that the MOG2 algorithm significantly outperforms its competitors, adapting efficiently to variations in the environment and maintaining a low number of false positives.

In terms of object classification, the YOLO-v5 model has demonstrated exceptional accuracy and speed, critical for real-time detection. Metrics derived from the confusion matrix, such as accuracy (the ratio of actual positive identifications to total predicted positive identifications), sensitivity (the model's ability to correctly identify true positives), and specificity (the ability to avoid false positives), indicated that our system effectively minimizes false alarms while maintaining a high rate of correct detection.

The overall evaluation of the system, combining both background subtraction and object classification, was performed through per-frame processing latency and alarm policy compliance. The results indicated that our system manages to maintain an average of 15 FPS and a per-frame latency of 0.54 seconds, thus ensuring real-time operation. The alarm policy, designed to differentiate between real and potential threats, was validated in diversified test scenarios, demonstrating the ability to reduce unneeded interruptions and focus the attention of security personnel on critical events.

D. Practical Applications and Use Cases

Our results have not only demonstrated the technical feasibility of an advanced surveillance system for real-time detection of people and objects on train tracks, but also highlighted its practical importance through experiments in real-life situations. Below, we present two scenarios that illustrate the system's ability to significantly improve railway safety:

Early Detection of Objects on the Track: In a hypothetical incident, our system detected a piece of heavy machinery inadvertently left on a curved section of track, where direct visibility is nil. The immediate detection of the object by our system allowed the control center to be alerted well in advance. This preventive action facilitated the halting of the approaching trains, avoiding a potential serious accident. This case highlights the importance of the system's ability to detect unexpected objects, especially in areas of difficult visibility.

Vandalism Prevention: In another scenario, a group of individuals was identified accessing the tracks at night with the intention of vandalizing parked cars. Early detection of the anomalous movement and accurate classification of 'persons' in the track area triggered security alarms. The rapid response of the security team, guided by the system's alerts, prevented the act of vandalism, thus ensuring both the safety and the security of the railcars.

VI. DISCUSSION AND COMPARISON

CV surveillance effectively automates and improves efficiency in the monitoring of the railroads.

The fusion of the two types of algorithms within our solution plays a vital role in maximizing real-time processing and energy efficiency. By optimizing the MOG2 background subtraction algorithm and the YOLOv5 detection technique to operate together seamlessly, the proposed system achieved a significant reduction in computational resource consumption. This optimization is reflected in the system's ability to process images in real time without compromising accuracy, even on equipment with limited capabilities. The complete source code and project documentation are publicly available on GitHub, allowing other researchers and developers to explore and contribute to the evolution of this railway surveillance system.

Our system uses an inexpensive regular camera per monitored area and object recognition to detect and analyze patterns in images and videos, enabling monitoring tasks to be performed with high

precision and a reduced number of false alarms. This is an operational innovation with respect to the current process performed by human operators. Furthermore, automated video surveillance is less error prone than human supervision, as it can operate continuously without any fatigue.

In this research, we have described a system for real-time detection of people and objects on the railroad using modern image processing, object classification and background subtraction techniques, with an architecture based on YOLO-v5 and the MOG2 algorithms, respectively.

Compared to other object detection algorithms, YOLOv5 has proven to have high accuracy in object detection in different types of images and videos. It also has a high processing speed, thanks to CUDA implementation, which makes it suitable for real-time applications, flexible and easy to use.

In our experiments, the MOG2 algorithm has generated enough mean time-per-frame processing speed to process the video continuously in real time in a regular PC, showing little artifacts or noise and a higher recovery by dynamically adapting to changes in the scene background, due to the adaptive Gaussian blending technique used. Furthermore, we have found that the MOG2 algorithm achieves the higher accuracy in identifying small objects and, according to the number of non-zero pixels in the experiments, more stability and speed in image changes than the other algorithms evaluated.

Regarding image classification, the accuracy metric for the "Person on track" event detection and the accuracy metric for the "Train object" event detection are more than acceptable for the correct operation of the system on a daily basis, as verified in the experiments, which means that the model is predicting correctly every time it detects those objects.

Our research contribution focuses on combining the most advanced current techniques, such as MOG2 background subtraction and YOLO-v5 neural networks. By leveraging the latest advancements in these algorithms, we aim to significantly enhance the effectiveness and efficiency compared to other similar systems.

S. Oh et al. [5] employ multiple cameras, combining frame difference information, thresholding, labeling, and fusion with laser sensor detection. In contrast, our approach differs by not requiring expensive components, utilizing only a single standard camera per platform.

On the other hand, the work of T. Xiao et al. [8] and A. D. Petrović et al. [15] describes various approaches to detecting obstacles on the rails. They propose a non-contact multisensory method using cameras and a LiDAR device, as well as a combination of deep neural networks and edge detection methods with cameras. Both proposals aim to detect obstacles on the tracks using on-board devices. In contrast, our research focuses on detecting objects and people on the platforms.

Finally, the work of H. Pan et al. [19] focuses on railroad signal detection and issues related to blocked track lines. Their approach differs from ours as it emphasizes track line detection and utilizes a combination of multi-task learning networks.

In general, our proposal is affordable and more flexible in generating alarms and pre-alarms. This flexibility allows for the easy implementation of new alarms within the surveillance logic. It is scalable and can launch multiple instances for each camera in a single PC.

In short, our solution prioritizes railroad safety, particularly the safety of people. Unlike other approaches, our system is specifically designed for object and person detection on the railroad. Additionally, our research shows that our solution can be executed in real time on a standard PC.

VII. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

The performance of these experiments, as presented in this study, includes several limitations that are crucial to contextualize the results obtained.

The effectiveness of artificial intelligence models is intrinsically linked to the diversity and quantity of data used for training. In this case, the data may be limited in terms of scenario variability, weather conditions and types of obstacles, which could affect the generalization of the model in real situations. These surveillance systems operate in a dynamic environment where conditions can change drastically, as in the case of lighting conditions. Such variations can affect detection accuracy, as the system may not have been exposed to all possible conditions during the training phase.

Effective implementation of surveillance systems in real-world environments requires seamless integration with existing railway control infrastructure. This aspect can present significant technical challenges, especially in older systems or those that are not designed to integrate with AI-based solutions. Furthermore, our ongoing research into the use of specialized hardware technologies, such as low-consumption FPGAs and ASICs, for future iterations of the system aims to substantially reduce electrical consumption. These efforts are not only in line with enhancing the operational efficiency of rail surveillance systems but also contribute significantly to the sustainability and environmental impact of rail transportation. The development of energy-efficient surveillance solutions will be a cornerstone of our future research efforts, ensuring that technological advancements in railway surveillance align with ecological responsibility.

In addition, it is crucial to maintain a careful balance between ensuring security and respecting individual privacy. The deployment of advanced surveillance technologies must be managed with careful consideration of ethical implications and the protection of personal freedoms.

Another major challenge is the ability of the system to efficiently monitor a large number of cameras simultaneously. Railway surveillance often requires an extensive network of cameras to cover all relevant areas. The increase in the number of cameras poses challenges in terms of real-time data processing and analysis, which can impact the speed and accuracy of detecting objects and people on the tracks.

The described research highlights the feasibility of implementing a real-time railroad track surveillance system using affordable technology and advanced image processing algorithms. The challenges faced range from variations in environmental conditions to the ability to differentiate between relevant and non-relevant objects. These limitations have been overcome by iterative adjustments to the algorithms, demonstrating the adaptability and robustness of our approach. However, we found certain limitations inherent to our study, particularly in terms of the availability of railway surveillance datasets for training, which may influence the generalization of the model to all possible practical situations. Future research will benefit from a larger and more diverse dataset, allowing a more robust generalization of the model. In addition, integration with existing railway control systems presents a fertile field for further exploration, which seeks a more holistic and automated implementation of real-time monitoring.

In the context of advances in rail surveillance, it is imperative to look to the future and identify key areas where further research can lead to significant improvements in the safety, efficiency, and sustainability of rail transportation. Despite progress in applying computer vision techniques and neural networks, there are substantial opportunities to expand and enrich this field. Future research could explore the

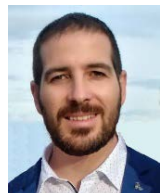
implementation of an intuitive and efficient user interface for railway surveillance systems. This includes customizable dashboards, real-time alerts, and advanced data visualizations that enhance monitoring and decision-making. Utilizing user-centered design techniques and incorporating operator feedback will be crucial in developing solutions that are not only technically advanced but also accessible and easy to use for surveillance personnel.

Another important area for future research is the integration of rail surveillance with urban transportation systems. This approach would allow for more efficient traffic management and improved coordination during emergencies or service disruptions. Research should focus on developing standardized and secure communication protocols that facilitate real-time information exchange between different modes of operation, such as passenger, maintenance, and freight. This data exchange could significantly improve emergency response, optimize resource allocation, and minimize service interruptions.

REFERENCES

- [1] S. Glimne, R. Brautaset and C. Österman, "Visual Fatigue During Control Room Work in Process Industries," vol. 65, no. 4, pp. 903-914, doi: 10.3233/WOR-203141. PMID: 32310219; PMCID: PMC7242839. 2020.
- [2] Y. Lei, T. Tian, B. Jiang, F. Qi, F. Jia, Q. Qu, "Research and Application of the Obstacle Avoidance System for High-Speed Railway Tunnel Lining Inspection Train Based on Integrated 3D LiDAR and 2D Camera Machine Vision Technology," *Applied Sciences*, vol. 13, no. 13, p. 7689, 2023.
- [3] M. Li, B. Peng, J. Liu and D. Zhai, "RBNet: An Ultrafast Rendering-Based Architecture for Railway Defect Segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1-8, 2023.
- [4] R. Goel, A. Sharma and R. Kapoor, "An Efficient Object and Railway Track Recognition in Thermal Images Using Deep Learning," *Emergent Converging Technologies and Biomedical Systems: Select Proceedings of ETBS 2021* (pp. 241-253). Springer Singapore. 2002.
- [5] S. Oh, S. Park and C. Lee, "A platform surveillance monitoring system using image processing for passenger safety in railway station," *ICCAS 2007 - International Conference on Control, Automation and Systems*, pp. 394-398, January 2007.
- [6] M. Arastounia, "Automated recognition of railroad infrastructure in rural areas from LIDAR data," *Remote Sensing*, vol. 7, no. 11, 2015.
- [7] B. Le Saux, A. Beaupère, A. Boulch, J. Brossard, A. Manier, and G. Villemin, "Railway detection: From Filtering to segmentation networks" *Proc. IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, Valencia, Spain, Jul. 2018.
- [8] T. Xiao, Y. Xu, and H. Yu, "Research on Obstacle Detection Method of Urban Rail Transit Based on Multisensor Technology," *Journal of Artificial Intelligence and Technology*, vol. 1, no. 1, pp. 61-67, 2021.
- [9] S. Taori, V. Kakade, S. Gandhi, and D. Jadhav, "Multi Sensor Obstacle Detection on Railway Tracks," *International Research Journal of Engineering and Technology*, vol. 8, no. 3, pp. 826-829, 2021.
- [10] F. Kaleli and Y. S. Akgul, "Vision-based railroad track extraction using dynamic programming," *Proceedings of 12th International IEEE Conference on Intelligent Transportation Systems*, St. Louis, MO, USA, oct. 2009, pp. 1-6.
- [11] Y. Wang, L. Wang, Y. H. Hu, and J. Qiu, "RailNet: A Segmentation Network for Railroad Detection," *IEEE Access*, vol. 7, pp. 143772-143779, 2019.
- [12] M. Ghorbanalivakili, J. Kang, G. Sohn, D. Beach and V. Marin, "TPE-Net: Track Point Extraction and Association Network for Rail Path Proposal Generation," 2023 IEEE 19th International Conference on Automation Science and Engineering (CASE), Auckland, New Zealand, 2023, pp. 1-7, doi: 10.1109/CASE56687.2023.10260541.
- [13] M. Qasim Gandapur, and E. Verdú, "ConvGRU-CNN: Spatiotemporal Deep Learning for Real-World Anomaly Detection in Video Surveillance System," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4, pp. 88-95, 2023.
- [14] M. Adimoolam, S. Mohan, 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, pp. 112-120, 2022.

- [15] A. D. Petrović, M. Banić, M. Simonović, D. Stamenković, A. Miltenović, G. Adamović, and D. Rangelov, "Integration of Computer Vision and Convolutional Neural Networks in the System for Detection of Rail Track and Signals on the Railway", *Applied Sciences*, vol. 12, no. 12, p. 6045, 2022.
- [16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," Jun. 2015, doi: 10.48550/arXiv.1506.02640.
- [17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2980-2988.
- [18] T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context", in: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) *Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science*, vol 8693, Springer, Cham, 2014, pp. 740-755.
- [19] H. Pan, Y. Li, H. Wang, and X. Tian, "Railway Obstacle Intrusion Detection Based on Convolution Neural Network Multitask Learning," *Electronics*, vol. 11, no. 17, p. 2697, 2022.
- [20] K. Matt Nolan, "L'intelligence artificielle au service de la sécurité sur les voies de la ligne Docklands Light Railway du métro londonien," *Mobilité Urbaine, France, Intelligence Artificielle, Sécurité. Mar. 2022*.
- [21] Z. Chen, J. Yang, C. Yang, "BrightsightNet: A lightweight progressive low-light image enhancement network and its application in "Rainbow" maglev train," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 10, p. 101814, 2023.
- [22] C. Meng, Z. Wang, L. Shi, Y. Gao, Y. Tao, and L. Wei, "SDRC-YOLO: A Novel Foreign Object Intrusion Detection Algorithm in Railway Scenarios," *Electronics*, vol. 12, no. 5, p. 1256, 2023.
- [23] Z. Chen, J. Yang, L. Chen, Z. Feng, L. Jia, "Efficient railway track region segmentation algorithm based on lightweight neural network and cross-fusion decoder," *Automation in Construction*, vol. 155, p. 105069, 2023.
- [24] Z. Feng, J. Yang, Z. Chen, and Z. Kang, "LRseg: An efficient railway region extraction method based on lightweight encoder and self-correcting decoder", *Expert Systems with Applications*, vol. 238, part F., p. 122386, 2024.
- [25] Z. Chen, J. Yang, Z. Feng, and H. Zhu, "RailFOD23: A dataset for foreign object detection on railroad transmission lines," *Scientific Data*, vol. 11, article no. 72, 2024.
- [26] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004*, vol. 2, Cambridge, UK, 2004, pp. 28-31, doi: 10.1109/ICPR.2004.1333992.
- [27] F. Porikli and O. Tuzel, "Bayesian background modeling for foreground detection," in *Proceeding of the ACM International Workshop on Video Surveillance and Sensor Networks (VSSN 2005)*, November 2005, pp. 55-58.
- [28] T. Cover and P. Hart, "Nearest neighbor pattern classification", *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27, 1967, doi: 10.1109/TIT.1967.1053964.
- [29] R. Kalsotra and S. Arora, "Background subtraction for moving object detection: explorations of recent developments and challenges," *The Visual Computer*, vol. 38, pp. 4151-4178, 2022. <https://doi.org/10.1007/s00371-021-02286-0>
- [30] B. Farnham, S. Tokyo, B. Boston, F. Sebastopol, and T. Beijing, "Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. Concepts, Tools, and Techniques to Build Intelligent Systems," O'Reilly Second Edition, 2017.



Dr. Domingo Martínez Núñez

Domingo Martínez Núñez is an Inspector at the Central Control Station of Metro de Madrid, S.A, currently serving as an Engineer at the Ticketing Development & Compliance Center. He has more than 10 years of professional experience. Domingo received professional training in electronics from the Army Polytechnic Institute in Zaragoza (I.P.E.) and holds a Degree in Computer

Engineering with a mention in Software Engineering at the International University of La Rioja (UNIR).



Dr. Fernando Carlos López Hernández

Fernando López is a full-time associate professor in the Applied Mathematics Department at Universidad Complutense de Madrid (UCM). His current research interests lie in image processing, computer vision, neural networks, dynamic systems, statistical machine learning and data-driven science. Before joining UCM, he was a full-time associate professor at Universidad Internacional de La Rioja (UNIR). He served as an education manager of the Doctorate Program in Computer Science, and in a Course of Robotics for Education. In addition, he lectured in the Computer Science degree covering subjects such as Statistics, Algebra, Discrete Mathematics, Algorithms Complexity, Image Processing, Signal Processing, Computer Graphics, Compilers.



Dr. J. Javier Rainer Granados

J. Javier Rainer. Director of the OTRI (Office for the Transfer of Research Results), and currently Director of the AENOR-UNIR Chair. Director of the Master in Quality Assessment and Certification Processes in Higher Education. PhD in Industrial Engineering from the Polytechnic University of Madrid (Spain). Master in Project Management, and University Expert in Management and Audit of Quality Systems. Personal Development Program in Artificial Intelligence. He has participated as a researcher and head of several R&D projects, has national and international publications, is a member of several international technical committees and has extensive experience in private companies in the telecommunications sector, where he has performed management and project management functions. In the field of university management, he has been Director of the Industrial Organization and Electronics Area of the School of Engineering and Technology of UNIR, and Deputy Director of Research of the same, and also the Director of the Degree in Industrial Organization. Research lines related to cognitive systems, decision making and learning.

Combating Misinformation and Polarization in the Corporate Sphere: Integrating Social, Technological and AI Strategies

Alberto Tejero¹, Galena Pisoni², Ziba Habibi Lashkari¹, Sergio Rios-Aguilar^{1*}

¹ Universidad Politécnica de Madrid (UPM), Madrid (Spain)

² York St. John University, York (United Kingdom)

* Corresponding author: sergio.rios@upm.es

Received 2 February 2024 | Accepted 18 March 2024 | Early Access 22 March 2024

unir
LA UNIVERSIDAD
EN INTERNET

ABSTRACT

In an era where misinformation and polarization present significant challenges, this research examines the root causes within social networks and assesses how corporations can use AI technologies for prompt detection. This research uses a dual approach: a "telephone game" with 225 participants from a Spanish university to study the spread of misinformation, and interviews with 15 experts from three French tech companies to investigate technological solutions. The findings indicate that almost one-third of participants inadvertently contribute to polarization, and around one-quarter propagated misinformation. The study also identifies the existing tools enhanced by AI and Machine Learning that effectively detect misinformation and polarization in corporate settings. This investigation provides crucial insights for practitioners to strengthen their strategies against misinformation and technical challenges and opportunities.

KEYWORDS

Business Strategies,
Information Systems,
Misinformation,
Polarization.

DOI: [10.9781/ijimai.2024.03.003](https://doi.org/10.9781/ijimai.2024.03.003)

I. INTRODUCTION

SOCIETY expects online businesses to maintain ethical and socially responsible. Assisting companies and practitioners in making informed decisions amid the rampant and unchecked spread of misinformation is challenging [1] [2]. In the world of social media individuals consume and produce information, with a single post potentially reaching vast audiences rapidly. Moreover, discerning the trustworthiness of information is difficult for individuals, who often receive it from their network and lack knowledge of the original source [3] [4]. The study addresses the phenomena of misinformation and polarization and their impact on companies' online presence including their products or services. We also explore user tendencies in a controlled experiment.

Online social networks enable global access to and rapid dissemination of information [5]. The content on these networks influences users' decisions and opinions on issues [6]. For many users, social media has become the primary source of information [7]. Nowadays, instead of traditional media, people rely more on social networks for information and the spread of information [5].

The problem of misinformation has garnered significant attention, and been a topic of concern for a long time [8] [9] [10]. Also, there has been focus on clarifying the elements that contribute the

misinformation spread, identifying the user groups who share it, and understanding how businesses can defend their brands against it.

This research addresses these gaps in the social computing landscape by pursuing two objectives: firstly to identify the factors that result in business-related misinformation and polarization, and secondly, to develop an automated system for such misinformation targeting businesses.

According to Guess and Lyons [11], misinformation consists of false or misleading messages presented as informative content, such as elite communication, online messages, advertisements, or published articles. Therefore, misinformation can be defined as a claim that contradicts or distorts common understandings of verifiable facts. On the other hand, disinformation is a subset of misinformation spread with the intent to deceive, distinguishing it from misinformation that may be shared without any intention of mislead.

Fig. 1 and Fig. 2 illustrate an increased interest in misinformation and disinformation at social (Google trends) and scientific (Web of Science) levels in recent years. However, an analysis of the scientific articles indicates that this interest focuses on the intentional spread of information [12] [8] [13] [14].

Polarization has been linked to differences in policy, known as issue polarization. Nowadays, it encompasses a wider range of effects such as effective polarization and distrust towards opposing views [15] [16] [17].

Please cite this article in press as: A. Tejero, G. Pisoni, Z. Habibi Lashkari, S. Rios-Aguilar. Combating Misinformation and Polarization in the Corporate Sphere: Integrating Social, Technological and AI Strategies, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 4, pp. 117-125, 2025, <http://dx.doi.org/10.9781/ijimai.2024.03.003>

Our beliefs -and environment- can bias our perceptions without our awareness which can result in unintentional misinformation. The bar chart shows information about a specific event, which can always have different interpretations, depending on the observer even if the data remains the same for everyone.

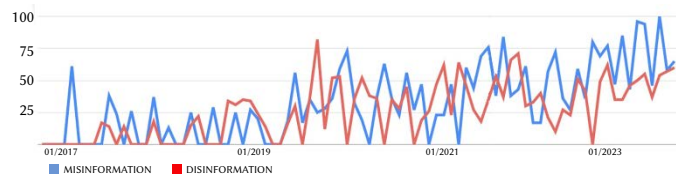


Fig 1. Analysis of the scientific publications of the concepts of Misinformation and Disinformation between the years 2017 to 2023 (Source: own elaboration based on Google Trends).

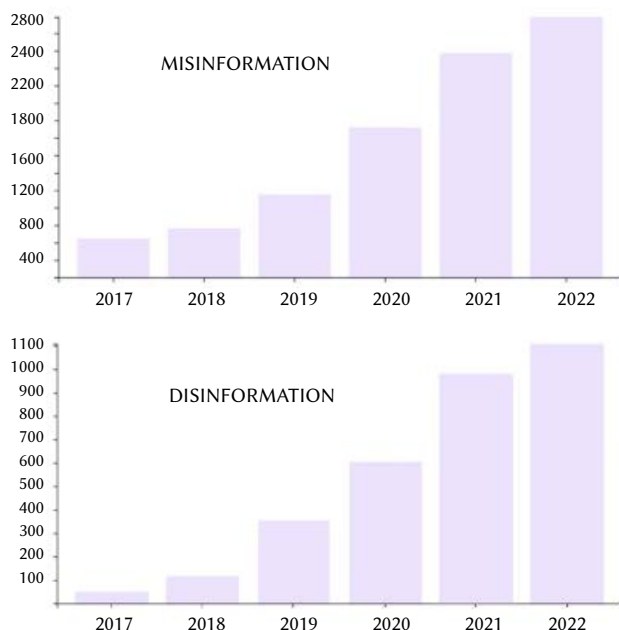


Fig. 2. Analysis of the scientific publications of the concepts of Misinformation and Disinformation between the years 2017 to 2022 (Source: own elaboration based on Web of Science).

This fact leads on many occasions to the generation of unintentional misinformation [16] [18]. That is why this study has been considered relevant in this work, to determine its level of importance and impact in relation to misinformation.

On the other hand, sociotechnical systems are intricate hybrids where human and technical resources collaborate to accomplish tasks [19] [20]. Performance depends on optimizing both the technical aspects -software and machinery infrastructure- and the social aspects such as, rules, procedures, roles and coordination.

Given the new demands of companies to constantly innovate and stay relevant, research in the domain of information systems (IS) concludes that IS are incomplete, always requiring enhancements. [21] [22]. As a result, the development of new technologies requires that sociotechnical constructions are revisited, and this drives the improvements developments of the IS of companies. These changes in the operations introduce new socio-technical constructions within the company.

The current trend towards digital transformation requires that companies maintain flexible and agile tech architecture that evolve with customer and organizational needs. Thus, modernizing information systems and technology in infrastructure is essential for

integrating diverse data and delivering digital services [23]. Therefore, a flexible underlying architecture is crucial and can accommodate integrations of new modules with different functions suited for the company's needs [24].

Architecture must also be robust enough to enable the company to develop new functionalities, instead of fixing bugs in the current system [25]. To achieve this, the technology architecture and design should be highly modular and loosely coupled, and support independent development of components and modules [26] [27]. The constantly evolving technological, organizational, and external environment means that business intelligence is being taken seriously in companies. Generally, these new services and component engineering cannot be provided without the insights gained through the application of Data Science and Artificial Intelligence (AI).

Artificial Intelligence has revolutionized how businesses utilize and manage data with Machine Learning (ML) -a subset of AI-, and automation, becoming ubiquitous in organizational operations. In recent years, AI has significantly helped companies in analysing and understanding their internal workings. AI models are on the rise, and while AI has contributed to misinformation [28], the social media literature has begun to investigate how AI can be used to predict and detect polarization and misinformation on social media [1] [29]. Still, literature has paid little attention to explaining how these models can be integrated into existing enterprise architectures, and how companies can leverage such available models. Another pressing issue is the lack of a clear definition of effective metrics for detecting misinformation and polarization, and what it means for a comment or opinion to be classified as polarized or misinformation [30]. Model accuracy, as applied as a metric in other fields, might not be relevant in this context. Despite its potential relevance to businesses, there appears to be a lack of research in this specific area to date.

Among others, Meta (formerly Facebook) is one of the companies that has utilized AI tools to identify and mitigate the spread of misinformation on their platform, implementing a range of policies and products [31]. These measures include adding warnings to content, reducing the distribution of questionable content, and removing harmful misinformation. They designed a specialized AI system to flag potential issues for review and to automatically detect new instances of previously identified misinformation, subsequently sending them to independent fact-checkers. They have developed technologies such as SimSearchNet++, which enhances image matching to identify variations of known misinformation images with high precision. Additionally, they have introduced new AI systems capable of detecting new variations of content that have already been debunked by fact-checkers, utilizing technologies like ObjectDNA and LASER cross-language sentence-level embedding.

This study investigates the phenomenon of users' information sharing about companies (or their products/services) on online social networking websites, and its implications for companies' technological developments. This research examines the factors that affects individuals' knowledge-sharing behaviours in social settings.

In this work, we face a set of research questions:

- What are the factors that lead to the polarization and degradation of messages, unintentionally generating misinformation? What are the typical user profiles that polarize and generate this misinformation?
- How companies can respond to misinformation and polarization presence regarding their products and services on social media? How can companies design / integrate components that can help them detect polarization and misinformation? How can companies integrate them with their existing information systems and enterprise architectures?

The gap that we try to fill in with our research is related to the analysis and identification of misinformation and polarization factors. Specifically, as previously anticipated, the focus of this work is on the analysis and identification of the factors related to non-deliberate misinformation.

We believe the finding will offer insights for enhancing corporate online strategies, leveraging AI-based tools for automation, and strengthening user trust- a vital aspect for sectors such as banking and healthcare.

We design a test to facilitate user knowledge exchange, employing a socio-technical approach. This allows us to explore human factors influencing misinformation and polarization, like gender and age, as well as the retention of information in friend circle. We also examine how companies can adapt their IS architectures to address these issues on social networks.

A mixed-method approach was employed, combining computational social science techniques with developer interviews and case studies. This aimed clarify the factors driving message polarization and degradation, intentional misinformation about companies, the profiles of those who generate such content, and how companies can technologically detect and prevent this spread on social media.

The research findings and contributions are further discussed from both aspects, sociological and technical. We propose a way forward for corporate (digital) social responsibility, and how they can implement responsible social presence on online platforms.

The structure of this document is as follows: section II outlines the research questions and objectives, section III describing materials and methods, section IV detailing the results, Section V discussing the results and Section VI presenting the conclusions.

II. MATERIALS AND METHODS

To explore our research topic, we will do experiments from two dimensions: technological and social.

This dual approach addresses gaps identified in the literature [14]: firstly, the factors which cause unintentional misinformation spread, and secondly, the absence of defined technological strategies for companies to manage online misinformation.

A. First Experiment: Social Dimension

The first experiment of this research was conducted at the empirical level within the social dimension, using the “Telephone game” method. Known globally by various names, including Chinese whispers, this game begins with one player whispering a sentence or phrase to the next person. This message is secretly passed along from person to person. When the secret has reached the entire group, the last person to hear it announces the secret aloud for everyone to hear. In general, by the game’s end, participants are surprised to hear how different the final version is from the version they heard. One of the main reasons for changes in the message is attributed to unintentional changes, such as impatience, wrong or faulty connections, although it can also be due to deliberate alterations on the part of the participants. The telephone game serves as a metaphor for the distortion that happens when information is passed along from person to person, whether first, second, third hand, or even more [32] [33] [34].

Specifically, this experiment aims to understand the factors causing misinformation and polarization through observation or direct experience with end users. Therefore, the focus of this experiment was on the users and their behaviour concerning misinformation and polarization, rather than the technology. For this reason, a low-tech test was designed, but one that would allow obtaining the maximum information about the users, who were the object of this phase of the research.

The designed test is described below (refer to Fig. 3 for more details):

- A text is selected, long enough (to offer the possibility of its transformation, summary, etc., at different levels), within a specific domain (news, politics, health, etc.).
- The original text is sent to a number N of people.
- Once they receive the text, each person must rewrite the text in their own words, creating a new version.
- After rewriting the text, each person must send the new text to a new group of people (M).
- After, each person must complete an anonymous form providing demographic information (age, gender, etc), which aids in group identification.
- People who receive the message must repeat steps 3 and 4 of the process to reach other people.
- The experiment concludes once a statistically sufficient sample is reached.
- Finally, the collected results obtained are analysed and categorized.

This experiment was designed using a simple technique, with the aim of facilitating participation, seeking to simplify the process as much as possible. In this way, the goal was to maximize the number of potential participants in the experiment in order to obtain a sufficiently significant sample for the research.

In this research, the topic of Covid-19 vaccines was used due to its significant global impact, making it a subject likely to draw heightened attention from participants. This can facilitate the identification of factors like misinformation and polarization. It’s important to note that the Covid-19 vaccine topic was not central to our research but was merely a means to conduct the study. The selected text comes from a real text identified on the social network X (formerly known as Twitter):

“The SAGE group indicated that no serious allergic reactions caused by the AstraZeneca vaccine have been recorded in the clinical trials against coronavirus. However, as for all vaccines, this should be administered under medical supervision, with appropriate medical treatment available in case of allergic reactions. In addition, anyone with an acute fever (body temperature over 38°C) should postpone vaccination until they are afebrile. However, the presence of a minor infection, such as a cold or low fever, should not delay vaccination”.

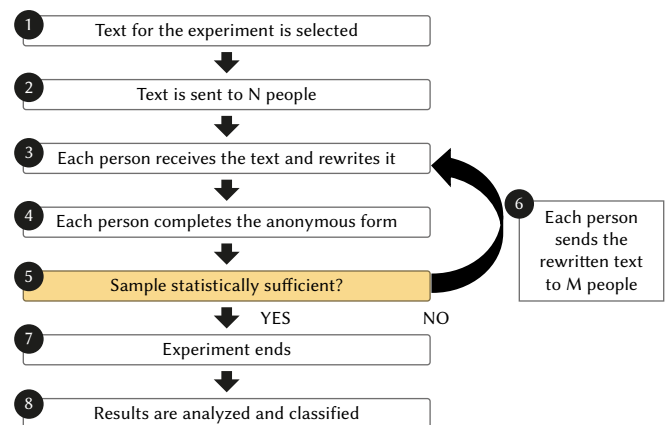


Fig. 3. Flowchart of the test designed for the experiment (Source: own elaboration).

The experiment was conducted within a university setting, encompassing students, faculty, and their close relatives, to ensure a diverse sample. To broaden the experiment’s scope and access the cultural dimension’s influence without adding undue complexity, it was executed in both English and Spanish.

To study misinformation, we identified specific keywords in the text to see how many keywords each user was able to retrieve and rewrite. The goal is to provide a clear view of how information is deteriorating regarding missing key values or, in this case, missing keywords.

For the polarization study, sentiment analysis was performed on each rewritten text, to observe how the polarity affects each user depending on the initial value of the original text. If the original text was identified with a positive voice, users who returned a negative polarity were analysed (and vice versa). Like this, it can be seen if people tend to have a positive or negative polarity based on their characteristics.

Equation (1) is used to obtain a statistically valid sample where N= Population size, z= Critical value of the normal distribution at the required confidence level, p= Sample proportion, e= Margin of error

$$\text{Sample Size} = \frac{\frac{z^2 * p(1-p)}{e^2}}{1 + \left(\frac{z^2 * p(1-p)}{e^2 N}\right)} \quad (1)$$

The original text was sent to 96 people (N) in English and Spanish, each of whom had to send the original text to three other people (M) to complete the experiment. In the end, 225 users participated in this experiment. In particular, 142 completed the experiment in Spanish and 73 in English. The total time it took to conduct the experiment, from when the first messages were sent to when the last responses were received, was two weeks.

In the experiment, N=96 people were initially surveyed (69 in Spanish and 27 in English). Each of the people in turn had to get M=3 other people to complete the experiment. To statistically guarantee the reliability of the samples, the following was done:

In Spanish:

- 276 (69 + 69 x 3) people is the maximum we can reach (size of the universe), while the people who have actually responded to the survey (sample) were: 142 people.
- When performing the calculations using (1), the value 136 is obtained. Therefore, 136 surveys were needed to ensure the reliability of the sample.
- Given that 142 people responded to the survey (142 > 136), the study was statistically feasible, with a 95% confidence level and a 6% margin of error.

In English:

- 108 (27 + 27 x 3) people is the maximum we can reach (size of the universe), while the people who have actually responded to the survey (sample) were: 73 people.
- The value 70 was obtained when performing the calculations with (1). Therefore, 70 surveys were needed to ensure the reliability of the sample.
- Given that 73 people responded to the survey (73 > 70), the study was statistically feasible, with a 95% confidence level and a 7% margin of error.

B. Second Experiment: Technological Dimension

This part of research focuses on the technological aspects and how companies can detect misinformation and polarization related to their products and services. We also explore how they can design or integrate technological components to analyse data from social networks and identify such issues.

The goal of this research segment is to assist companies in detecting disinformation in communications, as well as the factors that produce polarization, positive or negative, or misinformation (how, when, why, by whom). Analysing real time information will enable online

businesses to more effectively tackle misinformation and polarization on social platforms, allowing them to monitor and mitigate the factors that lead to disinformation and polarization in user message.

The interview guide was developed based on the conceptual frameworks [35] and [36]. Due to the exploratory nature of the study, the questionnaire consisted of open-ended questions. The interviews lasted between 30 minutes and 45 minutes. The open-ended questions guideline dealt with the following subjects:

- Current tech practices and the types of analysis companies perform on a daily basis regarding misinformation and polarization.
- How misinformation / polarization tools should be designed and chosen so to allow the online business and companies to be ethical, socially responsible, and protect customers.
- Integration of such solutions with the company's existing systems.

We followed a qualitative process to determine the collection and use of customer data from social media with company representatives coming from different industries. First, we obtained relevant documentation from the companies on the use of social media data and tools for their analysis. We relied on the data sources as specified in Table I. A descriptive review approach was employed to review relevant documentation on the enterprise architecture used in online business and digital companies and to identify the typical patterns in social media data analysis. The data for the case studies was collected in the Spring of 2023 and through interviews, we gained insights into the companies' internal processes from the respondents' perspectives.

TABLE I. DESCRIPTION OF COMPANIES INVOLVED AND PARTICIPANTS, WITH THEIR IDENTIFICATION NUMBERS MARKED WITH R, IN TECHNOLOGICAL EXPERIMENT (SOURCE: OWN ELABORATION)

Company	Domain	People involved in the interviews
Case company 1	Technical platforms realizations	2 (Strategic project manager, R1, Technical manager, R2)
Case company 2	Implementation of big data solutions for companies	2 (Technical manager, R3, Project Manager, R4)
Case company 3	Digital business consulting and services company that develops new digital services for clients, the main tasks: consulting and service design	3 (Technical coach, R5, R&D manager, R6, Business Developer, R7)

III. RESULTS

In this section we analyse the results we obtained from both experiments.

A. First Experiment Results

In the polarization analysis, the results obtained from the 215 participants of the experiment showed that 33.5% of the participants changed the polarization of the text they received. That is, approximately 1 in 3 people changed the polarization of the message they received without realizing it. In particular, in the experiment carried out in English, 37% of the participants changed the polarization of the text (27 out of 73), while in the experiment carried out in Spanish it was 31.7% (45 out of 142).

At the gender level, 34.71% of the female participants in the experiment changed the polarization of the original text they received, versus to 31.25% of the male participants. In particular, in the English version of the experiment, 38.3% of the female participants alerted the text's polarization compared to 32.43% in the Spanish version. In the case of men, 34.62% alerted the text in English experiment, while 30.88% did so in the Spanish experiment.

At the age level, the details of the participants changed polarization can be seen in Table II.

TABLE II. DETAIL OF PARTICIPANTS WHO CHANGED POLARIZATION AT AGE LEVEL (SOURCE: OWN ELABORATION)

Age range	Participants who changed polarization (%)	(English experiment) participants who changed the polarization	(Spanish experiment) participants who changed the polarization
18-30	35.16%	42.86%	30.38%
31-40	8.33%	14.29%	0%
41-50	61.54%	33.33%	14-29%
51-60	34.62%	25%	36.36%
+60	47.83%	0%	50%

During the misinformation analysis, a noteworthy 27% of the 215 participants revised their texts by incorporating half or less of the keywords present in the original material. That is, approximately 1 out of 4 people was unable to retain or maintain in their rewritten text more than half of the key words of the original text that they received. In particular, in the experiment carried out in English, 28.77% of the participants retained half or less of the keywords, while in the experiment in Spanish it was 26%.

At the gender level, 29.75% of women rewrote their texts using half or fewer of the original text's keyword, compared to 23.4% of men. In particular, in the experiment carried out in English, 36.17% of the participating women retained half or fewer words, whereas in the Spanish language experiment, the figure was 25.67%. For men, the English experiment saw 15.38% retaining half or fewer keywords, while in the Spanish experiment, this was higher at 25.68%.

Finally, at the age level, the details of the participants who rewrote their text using half or less of the keywords of the original text (%) can be seen in Table III.

TABLE III. DETAIL OF PARTICIPANTS WHO REWROTE THEIR TEXT USING HALF OR LESS OF THE KEYWORDS OF THE ORIGINAL TEXT (%) AT AGE LEVEL (SOURCE: OWN ELABORATION)

Age range	Participants who rewrote their texts using half or less of the keywords of the original text (%)	(English experiment) participants who rewrote their text using half or less of the keywords of the original text (%)	(Spanish experiment) participants who rewrote their text using half or less of the keywords of the original text (%)
18-30	24.22%	26.53%	22.78%
31-40	25%	28.57%	20%
41-50	23.08%	33.33%	14-29%
51-60	31%	50%	27.27%
+60	43.48%	0%	45.45%

B. Second Experiment Results

To implement new technological components and services for detecting misinformation and polarization, a company requires a holistic approach. This includes considering both its existing information systems and IT services, as well as the broader organizational context.

For this section, we used the findings from our case studies. We analysed and categorized the collected data to devise a generic architecture for the company's service that analyses social media data and detects misinformation and polarization.

This architecture is based on social media big data in any form (voice, text, image, video) and is analysed by a specific component. This component considers industry standards, company specific regulations and expert's insight in knowledge management. It also incorporates knowledge from the company's databases and knowledge bases through the existing information system and technology (See Fig. 4).

Companies utilize diverse data sources to identify misinformation and polarization, drawing from operational activities, internal records, and external references such as internet sources, audio recordings, and image files. This systematic data collection and analysis enables companies to derive insights and inform decision-making processes. Initially, unstructured data from various origins undergoes organization and cleansing, leading to the creation of a reliable knowledge base repository. This repository along with the factual database repository enables the dedicated component to predict and classify comments as misinformation or polarization. By integrating various types of data, a flexible data system is constructed that allows for ongoing refinement and utilization.

Fig. 4 illustrates the types of data utilized by companies for misinformation and polarization detection and application. Dashboards facilitate the visualization of the origins of misinformation and polarization, aiding in the identification of sources of polarized and inaccurate information. A repository of various AI tools is maintained for use within the component. Robust access control measures, including adherence to GDPR regulations, are employed to safeguard data privacy, confidentiality, and availability.

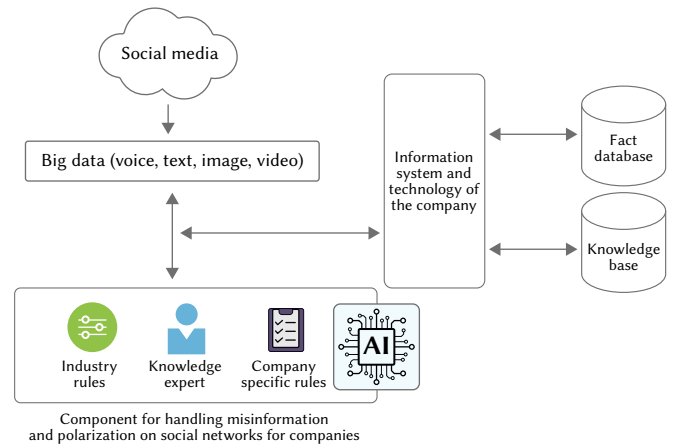


Fig. 4. Social networks misinformation and polarization component integration with existing technology for companies (Source: own elaboration).

Regarding the use of AI tools, the integration of ML techniques enhances operational resilience by improving the detection of misinformation and disinformation. Reference [37] presents a supervised ML model aimed at reducing operational interruptions, with a focus on supply chain processes. The model is praised for its role in strengthening decision-making through the assessment of diverse data sets. The study underlines the significant role of AI, and more specifically ML, in increasing the accuracy and reliability of false information identification, which is crucial for decision-making across various business specially in maintaining supply chains integrity.

However, integrating advanced AI-ML into existing systems presents significant challenges, requiring vast computational power and compatibility with established data systems [38]. The rapid evolution of AI technologies necessitates flexible and adaptable integration methods capable of managing intricate data analyses and accommodating ongoing updates. Achieving seamless integration involves not just technical alignment but also considerable investments in time, money, and specialized knowledge. Maintaining a

balance between introducing innovations and preserving operational consistency is essential for organizations to fully leverage the benefits of AI-ML enhancements.

A dedicated component like this would be of great importance to tackle misinformation spreading online, and could represent a key element of a system (observatory) to constantly monitor information flow in real time, enabling the issue of warnings about topics that require special caution. Respondents R2 and R5 recommend taking necessary measures to protect customer data according to existing customer protection laws. Therefore, this component should only utilize the latest research for detecting misinformation and polarization on social media, but also operate in compliance with General Data Protection Regulation (GDPR). Moreover, organizations must take appropriate technical and organizational measures to ensure the security of personal data. This includes implementing access controls, encryption, and other security protocols to prevent unauthorized access, use, or disclosure.

Respondents R2 and R5 commented that the current enterprise systems commonly deployed in companies on average are of flexible nature, often based on use of big data, and allow existing infrastructures to evolve and adapt new components. Currently, Big data is used for other business dimensions, such as customer profiling, cost of sales, market testing, price optimization, and understanding customer preferences. The use of social media data for detecting misinformation and polarization is less frequent, but according to R1, and R7, interest in this application is growing.

One challenge identified by the responders for the implementation of architecture as proposed above for companies, consists of assessing the level of complexity of integration, and companies need to be careful when choosing components that allow easy integration. Integration with existing architecture may be difficult, expensive and time consuming and the number of hours needed to have such integration in place need to be understood prior to enrolment in such project.

R1 said "Similar components have been selected in the past based on their functional requirements and ease of integration with the already existing solution".

Another challenge for businesses lies in assessing the risk of adopting a new component. Indeed, respondents acknowledged this, and the most important factors identified were cost for adoption, ease of customization, available support, licensees, as well long-term maintenance and reassessment if the component works fine. This is in line with previous research that supports such considerations for the adoption of new tech components for companies [39] [40].

R5 described the evaluation of risk like this "when the requirements are known, and the correct component identified, a risk analysis is undertaken to understand the long-term consequences in terms of support and maintenance of the component, and other risks, such as, commercial or cost related".

R6 also stressed the need for good documentation regarding the component: "Different components exist out there, however it is important to have a component that is well-documented and easy to maintain".

Some of the responders (R4 and R5) commented that when implementing such components, there must be a balance between speed to deliver such components, and maintaining the quality and stability of the overall architecture. R5 said "Solutions in which one supports deployment only of this particular component rather than deployment of the entire system must be put in place for scenarios as this one".

Respondents R3 and R6 said that they see it important that AI models are trained with company domain-specific knowledge, and not only as in previous literature, on misinformation and polarization

social media knowledge. For the moment, in previous literature [41] [42], approaches to improve misinformation and polarization rely on very basic deep learning models, not on company and domain specific knowledge, and there is a growing interest to expand knowledge on misinformation/polarization possibilities and understand how companies can run this kind of analysis on more complex deep learning models. At the same time, in the view of R3 and R6, AI models must be explainable; that is, in addition to detecting the sources of polarized and inaccurate information, they must explain to business users, how they have come up with the decisions taken, as well as why these comments and opinions are classified as misinformation and polarization compared to the others, to implement AI responsibly and transparently.

IV. DISCUSSION

Implications from research like ours are multiple. One relevant aspect, in discussions such as we start in this work, is the ethical and social responsibility of social platforms themselves [43] [44]. The social platforms are designed, so that more users use them, the more money they make. Social media companies have vested interest to keep users as much as possible on their platforms. They have a big number of personnel hired specifically to study what captures users' attention. As also development and deployment or changes in the current outlook are easy to perform, they launch a new feature of the social media platform in matter of weeks, so these companies have reached the point in which social media become irresistible for the users. We conclude with previous research, that social media initially were not created intentionally to spread misinformation and polarization, and that exactly these engagement mechanisms and vested interests of social companies, have contributed to this growing trend [45]

Perverseness of these technologies on one side, combined with the potential manipulability of these platforms as mentioned above, suggests that corporation's ethical and social responsibilities need to be revisited and include challenges beyond what was traditionally answered with corporate social responsibility agenda of companies. This is important to recognize, and, in this work, we aim to understand better this phenomenon, that is at least to understand the factors that led to it, and what companies can do in this respect [44].

Previous research has shown how a company approaches its corporate digital responsibilities varies from organization to organization, as well as the domain in which it operates [46] [47]. These norms and values that an organization follows are influenced by public opinion legal requirements, technological progress, industry factors, customer factors, and firm factors [46]. Previous research has also suggested that organizations need to follow ethical responsibility norms, at each step of data collection and use and make impact assessments [48].

We foresee positive correlation with companies' financial performance if companies implement responsible social media presence, our work in an initial first work in the domain, and we provide initial guidance to practitioners to drivers to misinformation and polarization as well as technological factors for social media misinformation and polarization mitigation on company level. We believe that future works should provide conceptual and analytical models that assist companies and managerial decision making, involving different stakeholders within the company; continued work in this field is timely and urgently needed.

Our findings lead us to implications and conclusions from both dimensions, social and technological. From the social dimension experiment, in view of the results obtained in the experiment carried out on 215 participants to analyse the levels of polarization and misinformation, it is observed that 33.5% of the participants changed

the polarization of the text they received (approximately, 1 in 3 people changed the polarization of the message they received), while 27% of the participants rewrote their texts using half or less of the keywords in the original text (approximately 1 out of 4 people was unable to retain or maintain in their rewritten text more than half of the key words of the original text that they received). This shows that polarization is a prevalent phenomenon, as a significant percentage of participants changed the polarization of the text they received. This highlights the influence of external factors on individuals' perspectives and the potential for polarization to occur. Regarding misinformation, this demonstrates that accurately retaining and conveying information is challenging and suggests a difficulty in understanding and conveying the intended message accurately.

Gender does not seem to be a determining factor in polarization, since, during the experiment carried out, women and men changed the polarization of the text they received at very similar levels (34.71% of women compared to 34.62% of men). However, a significant difference is identified between women and men in terms of misinformation. Specifically, women generated 29.75% compared to 23.4% of men, so gender does seem to be a factor to consider in the misinformation. Therefore, gender may play a role in misinformation, as there was a notable difference between women and men in terms of generating misinformation. Further exploration is needed to understand the underlying factors contributing to this disparity.

Considering age groups, a correlation is identified regarding misinformation across different age ranges, except for the 41-50 age group where it slightly decreases. Therefore, in general, an increase in misinformation levels is observed as age increases in the following ranges: 18-30 (24.22%), 31-40 (25%), 51-60 (31%), and +60 (43.48%). This could be related to the cognitive degeneration that occurs as we get older. However, in relation to polarization considering age, no correlations are observed, although there are very high levels of polarization in the age groups 41-50 (61.54%), followed by people aged +60 (47.83%).

It can be concluded, therefore, that both age and gender are determining factors in misinformation, whereas polarization only occurs partially. Knowing this fact, work could be done in the area of misinformation to improve levels, trying to act on gender and age factors. Strategies should focus on improving information retention and promoting critical thinking skills, particularly among older individuals.

From the technical dimension experiment, we noted that the volume of information (and misinformation) regarding companies that is available online is big, and the extent to which it's used poses challenges for business from misinformation and polarization points of view for the companies is significant. The businesses whose staff participated in this study have technologies and say that suited machine learning models can be easily integrated in existing systems to analyse data and understand perception of the company, as we propose in this work. However, it is noted that to have successful implementation, there must be a human oversight, and constant monitoring and integration of current customer protection rights acts, for responsible use of social media data. Further, there is an evolving understanding and attention regarding the importance of the issue, both within the society as whole and business, that even further accelerate the developments towards this topic.

What our interviewers told us is that the process of understanding long-term maintenance and viability of such projects within companies can be hindered due to conflicting agendas and priorities in the decision making of the company, yet change is inevitable, and it is to be expected that companies when deciding on such components will look for solutions less likely to change in future. Further

considerations for pace and scale of development of such solutions will need to be decided too.

Our research shows that companies' decision makers and suited personnel already have suited technical infrastructures, that easily can allow incorporation of such responsible misinformation and polarization detection components. We presented the requirements and design for a suited technology solution, and future work concerns the type, identification, assessment, and evaluation of suited ML models to use for this aim. The introduction of detection for use of misinformation and polarization has limitations in that i) the area of machine learning is only in its nascence, therefore detected results may be partially correct or sometimes maybe incorrect [41], and ii) the ethical implications for policies a company should adopt and how to behave in situation if misinformation or polarization is present and handled are not easy and straightforward task, requiring consolidation view by different stakeholders (not only the company) to develop a common understanding on how such behaviours are accounted. An even further challenge is such detection and decision making for such situations should be fully or partially automated.

As in any study, there are limitations. In this case, the social experiment was carried out starting from the university community and considering a small number of very general factors, such as age, sex, language, etc. Subsequent research where experiments will be carried out on a larger scale, or perhaps more focused on a set of specific factors to more thoroughly analyse their impact in terms of polarization and misinformation, could be of great interest to advance in this area. Finally, the text used in the present study was intended to be as neutral as possible, but in subsequent research it would be possible to work on different texts and levels of polarization. The generalizability of the findings in the technological experiment is limited in a number of ways as well. Our case studies were conducted with technological companies, that is providers of enterprise systems, which are the forerunners in implementing technology solutions for companies, and therefore are the first stakeholders willing to implement such solutions, yet such requests for solutions are expected to be on rise, due to hindering nature of the underlying problem.

Second limitation in the technological experiment comes from the qualitative nature of the study, with three technology companies interviewed, therefore future research will need to be performed on a larger scale. By highlighting the problem of presence of misinformation and polarization on social networks, especially in the reality of companies, this work shows on one side factors that lead to misinformation and polarization on social networks, and on another hand, what and how can companies approach this problem with specific technology tools developed for this aim. Future research considering different deployment models and training models can enhance the currently proposed solution. Our study of both, factors and technical solutions for misinformation and polarization for companies are therefore first and an important step, but more research is still needed.

V. CONCLUSIONS

Our work is a first study trying to understand social and technical factors with respect to misinformation and polarization for companies. We understood that one third of participants in our experiments polarizes texts on social media when needing to retell them, while older people (age 41 and above) spread misinformation more often compared to the other age groups and tend to polarize information more easily. Technological analysis says that recent research has presented relevant components for the detection of misinformation and polarization and such components are easy to develop and integrate with existing enterprise architectures of companies, with

factors like cost for adopting, ease of customization, support, licenses, as well long-term maintenance being important for the adoption.

Our work can help practitioners in companies to improve their understanding regarding misinformation and polarization, provides them information on what they can practically do to implement such services, and provides awareness for important aspects for setting more sophisticated services in future.

Up to our knowledge, our work is a first attempt to understand the effects of polarization and misinformation on social media for companies and what can companies do to cope with such presence. This is an emerging research area that requires further investigation.

This study opens several avenues for future research. In light of the constantly changing dynamics of misinformation and polarization, especially within social media environments, it is crucial to delve into the incorporation of sophisticated AI and ML strategies. These strategies should be capable of adjusting to the intricate subtleties of human interaction. Future studies could focus on developing more sophisticated models that account for the dynamic context of misinformation spread and the varying degrees of polarization. Additionally, investigating the long-term impact of misinformation and polarization mitigation strategies on organizational trust and consumer behaviour could provide valuable insights for both academia and industry.

As this field continues to grow, interdisciplinary approaches combining insights from social sciences, computer science, and organizational behaviour could enhance our understanding of misinformation's multifaceted impact. Moreover, with the increasing importance of ethical considerations in AI development and deployment, future research should also focus on the ethical frameworks that guide the use of AI in combating misinformation, ensuring that these technologies respect privacy, consent, and fairness.

REFERENCES

- [1] M.A. Al-Asadi, A. Mustafa, S. Tasdemir, "Using artificial intelligence against the phenomenon of fake news: a systematic literature review," in *Combating Fake News with Computational Intelligence Techniques. Studies in Computational Intelligence*, vol 1001, M. Lahby, A.S.K. Pathan, Y. Maleh, W.M.S. Yafooz, Eds. Springer, Cham. https://doi.org/10.1007/978-3-030-90087-8_2 2022, pp. 39-54.
- [2] H. Allcott, M. Gentzkow, C. YU, "Trends in the diffusion of misinformation on social media," *Research & Politics*, vol. 6, no 2, p. 2053168019848554, 2019.
- [3] V.L. Rubin, "Deception detection and rumor debunking for social media," in *The SAGE handbook of social media research methods*, L. Sloan, A. Quan-Haase, Eds. London: Sage, 2017. p. 342.
- [4] V.L. Rubin, N. Conroy, Y. Chen, S. Cornwell, "Fake news or truth? using satirical cues to detect potentially misleading news," in *Proceedings of the second workshop on computational approaches to deception detection*, T. Fornaciari, E. Fitzpatrick, J. Bachenko, Eds. 2016, pp. 7-17.
- [5] J. Boase, B. Wellman, "The internet and email aid users in maintaining their social networks and provide pathways to help when people face big decisions," *PEW internet and American life project*, 2006.
- [6] C. Kadushin, *Understanding social networks: Theories, concepts, and findings*, Oxford university press, 2012.
- [7] E. Bakshy, I. Rosenn, C. Marlow, L. Adamic, "The role of social networks in information diffusion," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 519-528.
- [8] P.W. Kraft, N. R. Davis, T. Davis, A. Heideman, J.T. Neumeyer, S.Y. Park, "Reliable Sources? Correcting Misinformation in Polarized Media Environments," *American Politics Research*, vol. 50, no 1, pp. 17-29, 2022.
- [9] P.N. Petratos, "Misinformation, disinformation, and fake news: Cyber risks to business," *Business Horizons*, vol. 64, no 6, pp. 763-774, 2021.
- [10] D. Spohr, "Fake news and ideological polarization: Filter bubbles and selective exposure on social media," *Business information review*, vol. 34, no. 3, pp. 150-160, 2017.
- [11] A.M. Guess, B.A. Lyons, "Misinformation, disinformation, and online propaganda," *Social media and democracy: The state of the field, prospects for reform*, N. Persily and J. A. Tucker, Eds. Cambridge: Cambridge University Press, 2020, pp. 10-33.
- [12] J. Shin, L. Jian, K. Driscoll, F. Bar, "The diffusion of misinformation on social media: Temporal pattern, message, and source," *Computers in Human Behavior*, vol. 83, pp. 278-287, 2018.
- [13] E.K. Vraga, L. Bode, "Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation," *Political Communication*, vol. 37, no. 1, pp. 136-144, 2020.
- [14] Y. Wang, M. McKee, A. Torbica, D. Stuckler, "Systematic literature review on the spread of health-related misinformation on social media," *Social science & medicine*, vol. 240, p. 112552, 2019.
- [15] P.M. Fernbach, L. Van Boven, "False polarization: Cognitive mechanisms and potential solutions," *Current opinion in Psychology*, vol. 43, pp. 1-6, 2022.
- [16] A. Bessi, F. Zollo, M. Del Vicario, A. Scala, G. Caldarelli, "Trend of narratives in the age of misinformation," *PloS one*, vol. 10, no. 8, p. e0134641, 2015.
- [17] M.H. Ribeiro, P.H. Calais, V.A.F. Almeida, "Everything I disagree with is# FakeNews: Correlating political polarization and spread of misinformation," *arXiv preprint arXiv:1706.05924*, 2017.
- [18] Y.K. Chang, I. Literat, C. Price, J.I. Eisman, J. Gardner, A. Chapman, A. Truss, "News literacy education in a polarized political climate: How games can teach youth to spot misinformation," *Harvard Kennedy School Misinformation Review*, vol. 1, no. 4, 2020.
- [19] R. Cooper, M. Foster, "Sociotechnical systems," *American Psychologist*, vol. 26, no 5, pp. 467-474, 1971.
- [20] G.H. Walker, N.A. Stanton, P.M. Salmon, D.P. Jenkins, "A review of sociotechnical systems theory: a classic concept for new command and control paradigms," *Theoretical issues in ergonomics science*, vol. 9, no. 6, pp. 479-499, 2008.
- [21] D. Tilson, K. Lyytinen, C. Sørensen, "Research commentary—Digital infrastructures: The missing IS research agenda," *Information systems research*, vol. 21, no. 4, pp. 748-759, 2010.
- [22] A. Tiwana, B. Konsynski, A.A. Bush, "Research commentary—Platform evolution: Coevolution of platform architecture, governance, and environmental dynamics," *Information systems research*, vol. 21, no. 4, pp. 675-687, 2010.
- [23] A. MacCormack, "Product-development practices that work: How Internet companies build software," *MIT Sloan Management Review*, vol. 42, no. 2, p. 75, 2001.
- [24] P. Rodríguez, A. Haghighatkhah, L.E. Lwakatare, S. Teppola, T. Suomalainen, et. al. "Continuous deployment of software intensive products and services: A systematic mapping study," *Journal of systems and software*, vol. 123, pp. 263-291, 2017.
- [25] H.H. Olsson, J.Bosch, H. Alahyari, "Towards R&D as innovation experiment systems: A framework for moving beyond agile software development," in *Proceedings of the IASTED*, pp. 798-805, 2013.
- [26] S. Bellomo, P. Kruchten, R.L. Nord, I. Ozkaya, "How to agilely architect an agile architecture," *Cutter IT Journal*, vol. 27, no. 2, pp. 12-17, 2014.
- [27] N. Brown, R.L. Nord, I. Ozkaya, M. Pais, "Analysis and management of architectural dependencies in iterative release planning," in *2011 Ninth Working IEEE/IFIP Conference on Software Architecture*. IEEE, 2011. pp. 103-112.
- [28] S. Monteith, T. Glenn, J.R. Geddes, P.C. Whybrow, E. Achtyes, M. Bauer, "Artificial intelligence and increasing misinformation," *The British Journal of Psychiatry*, vol. 224, no. 2, pp. 33-35, 2024.
- [29] S. Kreps, R.M. McCain, M. Brundage, "All the news that's fit to fabricate: AI-generated text as a tool of media misinformation," *Journal of experimental political science*, vol. 9, no. 1, pp. 104-117, 2022.
- [30] Q. Su, M. Wan, X. Liu, C.R. Huang, "Motivations, methods and metrics of misinformation detection: an NLP perspective," *Natural Language Processing Research*, vol. 1, no. 1-2, pp. 1-13, 2020.
- [31] Meta. "Here's how we're using AI to help detect misinformation". Accessed: Feb. 15, 2024. [Online]. Available: <https://ai.meta.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/>
- [32] E. Breck, C. Cardie, "Playing the telephone game: Determining the

hierarchical structure of perspective and speech expressions,” in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 2004, pp. 120-126.

- [33] F. Breithaupt, B. Li, “Fact vs. affect in the telephone game: All levels of surprise are retold with high accuracy, even independently of facts,” *Frontiers in psychology*, vol. 9, p. 375712, 2018.
- [34] E.J. Mulder, D. Snijders, “Playing the telephone game in a multilevel polity: On the implementation of e-government services for business in the EU,” *Government Information Quarterly*, vol. 39, no. 2, p. 101526, 2022.
- [35] G. Walsham, “Interpretive case studies in IS research: nature and method,” *European Journal of information systems*, vol. 4, no. 2, pp. 74-81, 1995.
- [36] R.K. Yin, “Designing case studies,” *Qualitative research methods*, vol. 5, no. 14, pp. 359-386, 2003.
- [37] P. Akhtar, A.M. Ghouri, H.U.R. Khan, M. Amin ul Haq, U. Awan et al. “Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions,” *Annals of Operations Research*, vol. 327, no. 2, pp. 633-657, 2023.
- [38] J. Tello, L. Subramaniam, “The magic behind turning data into profit,” *Deloitte*. Accessed: Feb. 15, 2024. [Online]. Available: <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/deloitte-analytics/us-ai-institute-challenges-of-using-artificial-intelligence-final.pdf>
- [39] S. Butler, J. Gamalielsson, B. Lundell, C. Brax, A. Mattsson et al. “Considerations and challenges for the adoption of open source components in software-intensive businesses,” *Journal of Systems and Software*, vol. 186, p. 111152, 2022.
- [40] A. Davila, G. Foster, “Management accounting systems adoption decisions: Evidence and performance implications from early-stage/startup companies,” *The Accounting Review*, vol. 80, no. 4, pp. 1039-1068, 2005.
- [41] D. Caled, M.J. Silva, “Digital media and misinformation: An outlook on multidisciplinary strategies against manipulation,” *Journal of Computational Social Science*, vol. 5, pp. 123-159, 2022.
- [42] D. Cavaliere, G. Fenza, V. Loia, F. Nota, “Emotion-Aware Monitoring of Users’ Reaction With a Multi-Perspective Analysis of Long- and Short-Term Topics on Twitter,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4, pp. 166-175, 2023, <https://doi.org/10.9781/ijimai.2023.02.003>
- [43] M. Sandoval, *From corporate to social media: Critical perspectives on corporate social responsibility in media and communication industries*, Routledge, 2014.
- [44] C. Stohl, M. Etter, S. Banghart, D.J. Woo, “Social media policies: Implications for contemporary notions of corporate social responsibility,” *Journal of business ethics*, vol. 142, pp. 413-436, 2017.
- [45] S. Vosoughi, D. Roy, S. Aral, “The spread of true and false news online,” *Science*, vol. 359, no. 6380, pp. 1146-1151, 2018.
- [46] L. Lobschat, B. Mueller, F. Eggers, L. Brandimarte, S. Diefenbach et al. “Corporate digital responsibility,” *Journal of Business Research*, vol. 122, pp. 875-888, 2021.
- [47] B. Mueller, “Corporate digital responsibility,” *Business & Information Systems Engineering*, vol. 64, no. 5, pp. 689-700, 2022.
- [48] L. Martín-Gómez, J. Pérez-Marcos, R. Cordero-Gutiérrez, D. H. De La Iglesia, “Promoting Social Media Dissemination of Digital Images Through CBR-Based Tag Recommendation,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 6, pp. 45-53, 2022. <https://doi.org/10.9781/ijimai.2022.09.002>



Alberto Tejero

Dr. Alberto Tejero holds an Master Degree in Business Administration - MBA (Universidad Complutense de Madrid), an MSc in Computer Science (specializing in Cybersecurity), and a PhD in Telecommunications Engineering from Universidad Politécnica de Madrid (UPM). He is an Associate Professor at UPM, Director of the Master in Digital Innovation (UPM and European

Institute of Innovation and Technology), former Deputy Director of the Transfer Office at UPM and an expert in international Technological Innovation Management, with a focus on entrepreneurship and innovation ecosystems. Alberto Tejero advises Foundations and Councils on innovation and education, and actively participates in international innovation and entrepreneurship projects in Europe, Latin America, and the USA.



Galena Pisoni

Dr. Galena Pisoni is a Lecturer (Assistant Professor) in Data Analytics at York Business School, York St John University, and a Visiting lecturer at IAE School of management, Université Côte d’Azur, Nice, France. Her research interests include enterprise information systems, data science, and knowledge management for companies of different sizes. She holds a Ph.D. in Computer Science with a focus on technology design and human-centered design of socio-technical systems. In her past she was a visiting researcher at center for Entrepreneurship, Strategy and Innovation management, University of Twente, Faculty of Industrial Design Engineering, Delft University of Technology, School of Architecture (Smart Spaces), KTH, and the Netherlands Organisation for applied scientific research (TNO), Groningen. She has an academic track record in business informatics and business intelligence, with publications in international conferences and journals in the broad areas of information systems, business and innovation.



Ziba Habibi Lashkari

Dr. Ziba Habibi Lashkari is an Associate Professor in Finance at Universidad Politécnica de Madrid. She has elevated her academic career to this distinguished position, showcasing her expertise and dedication to the field of FinTech. She currently serves as the Coordinator of the Master of FinTech program in EIT Digital and is also the Coordinator of the EIT Digital Summer School. She holds a Ph.D. in Economics and received the Ph.D. Extraordinary Award in 2019. She also is the coordinator of the “FinTech” summer school in Madrid and has been recognized for her innovative ideas in cybersecurity for fintech, earning a diploma for best startup ideas at the 18th ActúaUPM. A member of the “Decision Analysis and Statistics Group (DASG)”.



Sergio Rios Aguilar

Dr. Sergio Ríos holds a PhD in Computer Science (2021, URJC, Spain), a PhD in Economics and Business Sciences (2013, University of Granada, Spain), and a MSc in Telematics and Information Systems (URJC, 2008, Spain). He is an Associate Professor and Deputy Director of the Computer Science School at Universidad Politécnica de Madrid (UPM). He is a Lecturer in Emerging Technologies and Business Innovation, as well as Lecturer in I&E at the European Institute of Technology (EIT) MSc Programme. He was also the Director of a recognized Research Group on Mobility and User Experience (MUX) for 4 years. He has worked as a consultant in projects for several international companies like Samsung Electronics, Nokia, INTU Group, and others. He is a member of the European Global Navigation Satellite Systems Agency’s (GSA) Galileo Raw Measurement Taskforce. His current research interests include AI-powered Business Information Systems, Mobile presence and attendance systems and Indoor location-based Services.

Selecting the Appropriate User Experience Questionnaire and Guidance for Interpretation: the UEQ Family

Jessica Kollmorgen¹, Andreas Hinderks², Jörg Thomaschewski^{1*}

¹ University of Applied Sciences Emden/Leer (Germany)

² University of Applied Sciences and Arts Hannover (Germany)

* Corresponding author: joerg.thomaschewski@hs-emden-leer.de

Received 6 June 2024 | Accepted 15 July 2024 | Early Access 22 August 2024



ABSTRACT

Measuring the user experience (UX) of products, systems and services is individual depending on the research question. On the one hand, the user's goals and environment play a role in the subjective evaluation. On the other hand, different UX factors are relevant depending on the product. In this case, it is practical to have a questionnaire family as an aid, whose questionnaires are geared towards these different use cases. The User Experience Questionnaire (UEQ) family allows researchers and practitioners to choose the right tool for efficient UX measurement from three questionnaire versions. This article summarizes the UEQ, its short version (UEQ-S) and a modular framework (UEQ+) with overall 27 UX factors and purposes in over 30 different languages. In addition, specific instructions and assistance are provided for the statistical evaluation and interpretation of the questionnaire results. With the help of a key performance indicator (KPI), benchmarks and an importance-performance analysis (IPA), the realization of UX measurements is made easier for researchers and practitioners. To make it even more convenient to choose the right questionnaire from the UEQ family, influencing factors on the UX measurement and recommendations for action are given.

KEYWORDS

Agile Methods, UEQ, UEQ+, UEQ-S, User Experience, UX, UX Measurement, User Experience Questionnaire, UX Management.

DOI: 10.9781/ijimai.2024.08.005

I. INTRODUCTION

RESEARCHERS and practitioners have been dealing with the integration of actual user needs into the development of products, systems, and services for years to avoid late and often costly changes. Similar to agile development, it is necessary to continuously and frequently gather feedback from users. This forms the basis of the user experience research field, which focuses on planning, measuring, and evaluating user feedback. That is done not only to ensure efficiency and effectiveness but also to address specific emotions and subjective beliefs such as fun of use or aesthetics before, during and after the use [1].

In practice, the question arises of how to make these subjective impressions tangible in order to set concrete goals for improvements, or even the validation of achieved goals. Qualitative measurement methods offer one approach to do this, which involves among other things evaluating specific designs or individual cases for interpretive analysis. User interviews, for example, are well-suited for this qualitative analysis, generating some relevant insights with just a few participants. These measurement methods answer the "why" question, but can be time-consuming and not always generalizable to specific product or company goals [2].

In contrast, quantitative measurement methods provide another approach, using statistical data to analyze and identify concrete trends

[3]. These methods often serve as the basis for defining milestones in practice. However, acquiring large amounts of these quantitative data can be a lengthy process. Related work has shown that interpretations should not be made about populations based on small samples [4], [5], as they are more susceptible to disturbances and external influences. At the same time, the measurement method must not be too burdensome to allow for participation within a realistic timeframe, such as within the scope of daily life. A fundamental goal of user experience research is therefore to structure and simplify the capture of subjective perceptions.

An established measurement tool for quantitative research, minimizing the effort of recording and participating, is user experience questionnaires [6]. These enable the capture of subjective impressions, opinions, and evaluations from users regarding various aspects of the user experience. This makes it possible to generate large amounts of statistical data that can be further used for the analysis of products, systems, and services. User experience factors such as functionalities (e.g., easy to use, understandable) or sensations (e.g., entertaining) can be captured with an appropriate time investment from participants [7].

An established questionnaire family in this regard is the long-established User Experience Questionnaire (UEQ) [8], which has been continuously analyzed and expanded since 2008. This questionnaire defines various validated factors that measure specific aspects of

Please cite this article as:

J. Kollmorgen, A. Hinderks, J. Thomaschewski. Selecting the Appropriate User Experience Questionnaire and Guidance for Interpretation: the UEQ Family, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 4, pp. 126-139, 2025, <http://dx.doi.org/10.9781/ijimai.2024.08.005>

user experience. However, as explained, the goal of user experience research is to simplify the answering of these questionnaires as much as possible, which is why a short version of the questionnaire (UEQ-S) [7] exists since 2017. It should be noted that not all identified UX aspects are suitable or relevant for every industry or product, which is why a modular version of the questionnaire (UEQ+) [9] was published in 2019. This modular version can be assembled for individual use cases or research questions. This questionnaire family thus allows for a comprehensive quantification of subjective data.

In this article, we assist researchers and practitioners in selecting the appropriate questionnaires and user experience factors for the individual use cases and research questions. In addition, we provide specific assistance for the statistical analysis and interpretation of the questionnaire results. An overview of the UEQ Family including its components and their relations is shown in Fig. 1.

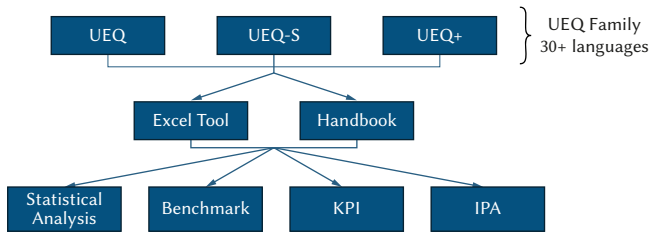


Fig. 1. Overview of the UEQ Family and its components, including the statistical analysis, benchmark, key performance indicator (KPI) and importance-performance analysis (IPA).

The structure of the article is therefore as follows: after an overview in Section II of basic concepts and related work on measuring the expected and perceived user experience with the help of questionnaires, Section III will first present the specific foundations and differences of the individual questionnaires of the UEQ family, including their measured factors and applicability, also in the context of other questionnaires. Subsequently, in Section IV, an insight into the interpretation quantitative UX measurement will be provided to researchers and practitioners on the example of the UEQ family, including relevant statistical analyses using methods such as the UEQ Tool, benchmarks, relevant key performance indicators, and the importance-performance analysis. These reference points and comparison values help to interpret and practically classify the measured results. Section V will focus on relevant user- and product-related influencing factors that can affect both the subjective perceived user experience and the interpretations of the results, which should be considered. Finally, this forms the basis for concrete recommendations for action and an outlook in Section VI.

II. BASIC CONCEPTS

This section introduces and explains some basic concepts that are necessary for the general argumentation and interpretation of this article, including the differentiation between usability and user experience and the measurement of user experience using questionnaires.

A. Usability and User Experience

If we imagine products, systems or services, such as online banking, then the users' expected requirements focus on tasks that they want to achieve with them. Factors such as efficiency and effectiveness are relevant for operation. This is associated with the concept of usability, which according to the established ISO 9241-11 is described as "the effectiveness, efficiency and satisfaction with which specified users achieve specified goals in particular environments." [10]

However, if we now take a look at modern social networks, factors have less to do with the fulfillment of tasks and more to do with individual perceptions, such as aesthetics or novelty. The concept of user experience (UX) thus expands the original usability concept with additional factors in order to provide a more holistic overview of the interaction between user and product¹. User experience can therefore be described as the "user's perceptions and responses that result from the use and/or anticipated use of a system, product or service" [11], including for example also emotions or beliefs.

Since the 1990s, there have been various models that attempt to describe usability and user experience as a whole. Each model has a different focus for mapping UX with the help of factors [1]. For example, Hassenzahl's [12] model takes a holistic approach to UX, which describes that UX consists of different dimensions: *pragmatic* and *hedonic* quality. While the *pragmatic* UX aspects relate to the usability and functionality of a product in a psychological context, the *hedonic* UX aspects relate more to emotions, such as the experience of pleasure or aesthetics. This theoretical construct also has a practical background, as could be also shown in a study from 2023 [13]. Here, it was shown that products with a differently weighted pragmatic or hedonic focus were also rated differently.

However, as the subjectivity of the aspects already makes clear, different users or target groups may judge the same product differently in terms of its UX. This may be due to the fact that they have different needs or abilities when using the product. For example, older users might experience social networks quite differently to younger, more tech-savvy users [14]. According to Hassenzahl (2003) [12], the different usage modes also play a role in the subjectivity of the evaluation. While the *goal mode* is aimed at the pure fulfillment of goals, in *action mode* the goal is not fixed in advance, but arises during use. In the latter case, the use of the product itself can also be the goal, e.g. arising from boredom.

It is therefore necessary to check the expected and perceived UX of products not just at one point in time, but continuously. This can refer to the time before (anticipated UX), during (momentary UX), after the use (episodic UX) and to the course over time (cumulative UX) [15]. To be able to examine this, the UX must first be made measurable [1].

B. UX Measurement

Users expect a high level of satisfaction when interacting with the product, both for simple and complex products that specialize in functionality rather than user satisfaction. They are expected to be able to use the product efficiently without much effort [1]. But how exactly can individual and subjective UX be continuously measured and how can generalized conclusions be drawn?

There are various evaluation approaches to prevent, for example, the design from mismatching with the user's mental model [16]. Well-known methods for this are, for example, user interviews or heuristic evaluation [17]. Relevant findings of these methods can ultimately be drawn to improve the UX of products. In order to implement changes in a standardized way without losing sight of other aspects that are already good, researchers and experts developed so-called usability and UX heuristics. These compromise guidelines for identifying and eliminating potential UX weaknesses at an early stage [18].

While these qualitative methods can also capture complex views, they are time-consuming and resource-intensive. They reach their limits when it comes to the scalability of UX measurement for many users and different target groups [14]. For example, it should be avoided that certain user groups, such as older or less educated people or people with disabilities, are excluded from use because they are not considered and thus experience a poor user experience [19].

¹ In the rest of the article, "product" is used as a substitute for "product, system or service".

In this case, quantitative user experience questionnaires are a suitable tool. Thanks to their simple scalability and comparability as well as efficiency and standardization in data collection, they allow a wide range of aspects to be surveyed and flexibly adapted to the target group without generating high costs. They can measure the expected and perceived UX of products and form a basis for improvements to the product under investigation [7].

C. UX Questionnaires

In recent decades, various standardized questionnaires have been established that measure both usability and user experience factors. A factor is equated with an actual and real UX aspect, and consists of various individual items that are assigned to this factor in advance using suitable methods such as statistical factor analysis. A standardized questionnaire therefore contains multiple UX factors, which in turn contain different items [1].

Various forms of items can be used when dealing with questionnaires. Commonly used are 5- or 7-point Likert scales, which users can use to express their level of (dis)agreement in relation to a short statement [20]. One example of this is the SUS (System Usability Scale) questionnaire, which is shown in Fig. 2.

1. I think that I would like to use this system frequently

Strongly disagree					Strongly agree
1	2	3	4	5	

Fig. 2. Example of an item from the System Usability Scale (SUS) [21].

However, the disadvantage of this item format is the scope for interpretation. Therefore, there are also semantic differentials that use certain clear, contradictory terms in relation to a UX factor. This is the case with the UEQ (User Experience Questionnaire), an example of which is shown in Fig. 3.

Please assess the product now by ticking one circle per line.

attractive	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	unattractive
------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	--------------

Fig. 3. Example of an item from the User Experience Questionnaire (UEQ) [22].

These contrasting attributes can reduce misunderstandings and inconsistencies in interpretation [20].

With the help of holistic questionnaires, it is therefore possible to determine the satisfaction of various relevant user groups with individual products and use cases. However, UX questionnaires differ in their area of application and focus. While the AttrakDiff [23], for example, lays a greater emphasis on hedonic than on pragmatic quality aspects, which can be not entirely suitable for the assessment of professional software [8], the UEQ [8] and the mCUE [24] are broader in scope and measure both the usability and the user experience for products in a wide range of application areas. Based on their regular use, it is possible to continuously record the expectations and perceptions of users and improve the quality of the UX on this basis [1].

III. FUNDAMENTALS

As explained in Section II, UX questionnaires enable a quick but also comprehensive measurement and evaluation of the expected and perceived UX of products for larger user target groups. For example, users can be presented with a short questionnaire when they leave a web service to enable data collection from large samples even with

little effort [14]. These questionnaires can be provided both online and in printed form.

Different questionnaires take various UX factors into account and are effective for their specific contexts. The UEQ Family is also designed for different contexts and offers a flexible approach. It is designed to be highly adaptable, allowing for customization to suit a wide range of use cases in UX measurement. This adaptability ensures the provision of relevant and accurate insights across different contexts and user groups.

The individual products of the UEQ family, the UEQ, UEQ-S (short version) and UEQ+ (modular version), are presented below with the respective descriptions of UX factors, items and intended use. The UEQ family is currently available in over 30 languages.

A. UEQ²

The User Experience Questionnaire was published in 2008 by Laugwitz et al. [8] to quantitatively measure the perceived UX of products with their respective functions. The UEQ follows three main objectives. Firstly, the application and analysis of the questionnaire should enable a **quick assessment** as far as possible. Secondly, with the help of end users, a **comprehensive impression of user experience** of the product under investigation should be gained. Thirdly, users should be given the opportunity to express feelings/impressions that arise when experiencing the product in a **simple and immediate** way as possible. In 2008, other available questionnaires also focussed on one or two main objectives, but in no case all three [8].

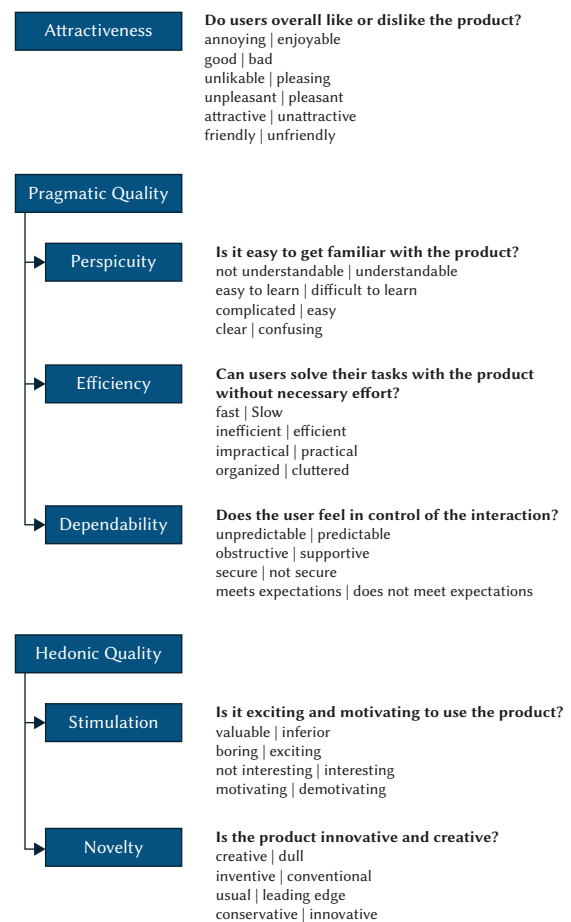


Fig. 4. Overview of the UEQ and the included UX factors with their meaning and corresponding items.

² available online free of charge under <https://www.ueq-online.org/>

The UEQ measures a total of six UX factors (see Fig. 4), which can be assigned to *pragmatic* and *hedonic* quality aspects as well as the attractiveness of the product. Each of the factors is measured using specific items. While the factor *attractiveness* is measured with six items, the other five factors are determined using four items each. As a result, the UEQ with the six UX factors contains a total of 26 items, which take a total of 3-5 minutes to answer [7].

The process of selecting the items was extensive [8]. In total, a set of originally more than 200 potential UX items was created in a study with 153 participants. In the next step, this selection was shortened with a number of experts to a rough version of 80 items. A statistical factorial analysis was used to check which of these items could be assigned to which of the six factors as clearly as possible and free from misinterpretation. After completion of the UEQ, the questionnaire was additionally validated in two studies [8].

Semantic differentials were chosen as the item format, which are recorded on a 7-point Likert scale with a range from -3 (worst) to +3 (best). Half of the positive attributes are on the right-hand side, the other half on the left-hand side of the questionnaire. This is to avoid the introduction of a response tendency with a one-sided item polarity [8], [25]. The distribution of the item order is randomized [7], [8]. An excerpt of the different items can be seen using the example of the factor *Efficiency* in Fig. 5.

To achieve my goals, I consider the product as

slow	○	○	○	○	○	○	fast
inefficient	○	○	○	○	○	○	efficient
impractical	○	○	○	○	○	○	practical
cluttered	○	○	○	○	○	○	organized

Fig. 5. Items of the factor *Efficiency* measured with the UEQ [22].

In 2019, the UEQ was expanded to include a key performance indicator (KPI) [26]. This was due to the desire of decision-makers to interpret single numerical values. To obtain an UX KPI, it was first necessary to measure the relative importance of the individual six UEQ factors in relation to the product under investigation. This means that an additional question on the importance of the specific factor was added to each of the six factors, which can be rated on a 7-point scale. An example of the factor *Efficiency* is shown in Fig. 6.

I can perform my tasks with the product fast, efficient and in a pragmatic way. The user interface looks organized.

Not important at all ○ ○ ○ ○ ○ ○ ○ Very important

Fig. 6. Item related to the key performance indicator (KPI) of the factor *Efficiency* measured with the UEQ [22].

These importance ratings are used to determine additional UX information (see Section IV.C).

At the beginning of further development (2008), the UEQ was originally available in German and English. However, in order to enable the most international use possible in a wide range of industries and use cases, the UEQ is now available for download in 36 languages, including Spanish, Finnish, Russian, Chinese or Turkish [27]. It is also available in a simpler version with simplified language, e.g. for children or people with disabilities [28]. Over time, the versions created have been repeatedly checked and improved, further reducing the risk of misinterpretation [29].

In order to avoid misunderstandings when implementing and evaluating the UEQ, a website [27], a handbook [22] and a supporting

Excel tool [30] are available to facilitate statistical analysis, so only the individual data need to be inserted.

This makes the UEQ an easy-to-use, reliable and valid method for quickly capturing the user experience of products within a few minutes, and is short enough to be completed both in printed and online form, even if demographic questions are added at the beginning or end of the questionnaire to gain an additional impression of the target group [14]. It can also supplement or support data from other evaluation methods with its subjective quality ratings [7].

B. UEQ-S³

Even if it only takes 3-5 minutes to complete the UEQ, there are situations in which this time is not available, for example if the user to be surveyed leaves a web service. Other possible situations would be that the respondents are to evaluate several product (variants) in one go, or the questionnaire is to be integrated into existing product experience questionnaires, which would exceed the reasonable total length. In these and other cases, it is helpful to be able to fall back on a short version.

For this reason, a short version of the UEQ, the UEQ-Short (UEQ-S), was published in 2017 by Schrepp et al. [7]. In the UEQ-S, the measurement of the six single dimensions like *Efficiency* were skipped, and only the meta-dimensions *Pragmatic Quality* and *Hedonic Quality* are measured. Instead of 26 items of the UEQ, these two dimensions are measured with a total of only 8 items, which are shown in Fig. 7 (4 items of the *pragmatic* factors as well as 4 items of the *hedonic* factors which fit best from the UEQ). The mean value of the 8 items is evaluated as the overall UX value [7].

The positive attributes of the semantic differentials are on the same side, as the focus of the short version is on the required completion time and it was assumed that a one-sided item polarity reduces the completion time of the questionnaire. Furthermore, the order of the items is not randomized: the first 4 items reflect the *pragmatic* quality, the other 4 items, the *hedonic* quality.

obstructive	○	○	○	○	○	○	supportive
complicated	○	○	○	○	○	○	easy
inefficient	○	○	○	○	○	○	efficient
confusing	○	○	○	○	○	○	clear
boring	○	○	○	○	○	○	exciting
not interesting	○	○	○	○	○	○	interesting
conventional	○	○	○	○	○	○	inventive
usual	○	○	○	○	○	○	leading edge

Fig. 7. Items of the short version of the UEQ (UEQ-S) [27].

A data set with 1867 data records was used as the basis for creating the UEQ-S, in which a total of 21 different products from different areas of application were assessed (e.g. webshops or business software) [7]. Based on this data, a main component analysis was carried out on all items of the *pragmatic* factors and then on the items of the *hedonic* factors. In order to prove that the scale is precise and that items do not overlap with other items, the 8 items with the lowest factor loadings according to the analysis were selected for the short version. The constructed UEQ-S was also evaluated in a further study with 47 people and 3 products. A high level of consistency was shown and the previously measured results were confirmed, meaning that the questionnaire is consistent and stable in practice [7].

³ available online free of charge under <https://www.ueq-online.org/>

The mean values of the four items were then compared with the mean values of the respective 16 and 12 items of the *pragmatic* and *hedonic* factors. It was found that the differences were close to zero, meaning that the short version is able to accurately predict the values of the full UEQ. As the UEQ-S is therefore a subset of the UEQ, this also means that the UEQ-S can be used in all 30+ languages in which the UEQ is available [7]. The short version provides an approximate assessment of the UX quality of a product based on higher level meta-dimensions and is not intended to replace the UEQ. It is therefore recommended to measure the six UX factors in detail if the results are to be interpreted for improvement potential. In such cases, e.g., when used together with usability tests, the small increase in efficiency in the completion time does not compensate for the loss of detailed impressions of the quality aspects.

C. UEQ+⁴

As the user experience is a subjective construct of various factors, it follows that not all of these UX quality aspects are actually relevant depending on the product and use case. Experience with the UEQ has shown that UX factors are missing (e.g. Trust [31]) and factors available in the UEQ are sometimes not needed (e.g. Stimulation [32]).

These considerations gave rise to the UEQ+ (UEQplus) questionnaire, which was published in 2019 by Schrepp and Thomaschewski [9]. This is a modular framework that allows researchers and practitioners to select the relevant UX factors for the desired use case from a catalog of UX factors.

Since the factors can be combined individually depending on the use of the questionnaire, it was necessary to change the randomized format of the semantic differentials to a one-sided item polarity in which all positive terms are positioned on the right-hand side. Since the order of the selected factors should also be able to be determined individually, all items of a factor were grouped and placed in context to prevent misinterpretation. For this purpose, a sentence was added above the items of a factor (see Fig. 8). However, the 7-point Likert scales were retained as their use in the UEQ has been empirically validated. All factors therefore have the same response format so that they can be easily combined [9].

In order to be able to make a judgment as to whether the selected factors are actually relevant for the use case or the product, an importance rating was also added to each UX factor using a 7-point Likert scale (see Fig. 8), similar to the KPI extension of the UEQ.

To achieve my goals, I consider the product as						
slow	○	○	○	○	○	fast
inefficient	○	○	○	○	○	efficient
impractical	○	○	○	○	○	practical
cluttered	○	○	○	○	○	organized
I consider the product property described by these terms as						
Completely irrelevant	○	○	○	○	○	Very important

Fig. 8. Example of the UX factor *Efficiency* and the corresponding items [5].

To add further factors to the six UX factors of the UEQ, an empirical study was conducted in 2018 with 192 participants, the results of which were examined in a principal component analysis. Based on this, four items with the highest factor loadings per factor were used for newly created factors. The study resulted in the following UX factors, which

can be measured using the UEQ+ [9], consisting of the factors of the UEQ as well as new factors:

- **Attractiveness:** Overall impression concerning the product. Do users like or dislike it?⁵
- **Perspicuity:** Impression that it is easy to learn how to use the product.
- **Efficiency:** Impression that tasks can be finished without unnecessary effort.
- **Dependability:** Impression to be in control of the interaction with the product.
- **Stimulation:** Impression that it is interesting and fun to use the product.
- **Novelty:** Impression that the product design or product idea is creative and original. [33]
- **Trust:** Subjective impression that the data entered into the product are in safe hands and are not used to the detriment of the user.
- **Aesthetics:** Impression that the product looks nice and appealing.
- **Adaptability:** Subjective impression that the product can be easily adapted to personal preferences or personal working styles.
- **Usefulness:** Subjective impression that using the product brings advantages, saves time or improves personal productivity.
- **Intuitive Use:** Subjective impression that the product can be used immediately without any training, instructions or help from other persons.
- **Value:** Subjective impression that the product is of high quality and professionally designed.
- **Trustworthiness of Content:** Subjective impression that the information provided by the product is reliable and accurate.
- **Quality of Content:** Subjective impression that the information provided by the product is up to date, well-prepared and interesting. [9]

Factors that are specifically geared towards household appliances (e.g. washing machines) were also included:

- **Haptics:** Subjective feelings resulting from touching the product.
- **Acoustics:** Subjective impression concerning the sound or operating noise of the product. [9]

These factors with the associated items were then evaluated in a further study using three product categories (webshops, video platforms, programming environments) and two products each. A high factor quality was proven in this study. In addition, the selected factors for each product were considered important for the participants [9].

In further studies after 2019, additional UX factors were also included in the factor catalog. For example, the UEQ+ can now measure UX factors of voice assistants (e.g. Siri or Alexa):

- **Response behavior:** Impression that the voice assistant behaves respectful and trustworthy.
- **Response quality:** Impression that the responses of a voice assistant cover the user's information needs.
- **Comprehensibility:** Impression that the voice assistant correctly understands the users instructions and questions using natural language. [34]

The UX of complex medical devices (e.g. MRI and CT scanners) can also be measured:

- **Result quality:** Can goals and results be fully and accurately achieved by using the product?

⁴ available online free of charge under <https://ueqplus.ueq-research.org/>

⁵ In contrast to the UEQ, the factor *Attractiveness* is measured with 4 items in the UEQ+.

- **Hardware security:** Does the hardware bear risks, which might be hazardous to health?
- **Risk handling:** Can users identify and handle risks and errors? [35]

And factors for the influence of using or owning a product on social connections/status can also be measured:

- **Identification:** Impression that using or owning a product influences the social status.
- **Social interaction:** Impression of the user that the product supports social activities or helps to build social contacts.
- **Social stimulation:** Impression of the user regarding the anticipated social gains resulting from their interaction with a product.
- **Social acceptance:** Impression of the user regarding how they are accepted and approved by others and themselves when using a product. [36]

Over the years, the UEQ+ has been developed as a modular questionnaire that allows researchers and practitioners to measure a total of 27 UX factors depending on the research question. The factors, which all have the same item and response format, can be combined individually and are easy to answer.

However, there are also some additional expenses compared to the original UEQ. The effort required to compile and evaluate the modular questionnaire is higher, as it is not always known which factors need to be selected for each product. However, we will present impact factors and recommendations for action below to make the choice easier. In addition, a handbook [22] and a supporting Excel tool [37] are also offered for the UEQ+, to facilitate statistical analysis, so only the individual data need to be inserted. Benchmarks are also provided, for example, to make it easier to interpret the results. The UEQ+ is also currently available in 25 languages [33].

D. Differentiation From Other Established Questionnaires

As already explained in Section II, there are a large number of UX factors and corresponding questionnaires that measure these factors. This section therefore aims to show the relationship between the UEQ family and established related questionnaires. For more in-depth analyses, corresponding studies are recommended [1], [20].

In a comparison of the German-language version of the UEQ from 2018 [38] with the German-language questionnaires VISAWI (Visual Aesthetics of Website Inventory) and meCUE, three products with different usage contexts were selected. While the VISAWI measures factors such as *Variety* or *Colorfulness*, the meCUE measures three modules based on the “Components of User Experience” model: Product perception (factors e.g. *Usefulness*, *Usability* or *Visual Aesthetics*), user emotions (positive and negative) and consequences (factors *Product loyalty* or *Intention to use*). On the one hand, there were high correlations, e.g. between the *hedonic* UEQ factors and the VISAWI factors, which was to be expected according to the underlying psychological model of the questionnaires. On the other hand, however, the correlations of the questionnaire factors varied depending on the product, which means that other influences such as the context of use may be present here.

In another comparison of the questionnaire results of the UEQ-S with SUS and UMUX-LITE, a study [13] with four products showed that all three questionnaires delivered almost identical results in the evaluation of the UX quality of the products. This was also evident in the scale correlations, as UMUX-LITE and SUS showed high correlations with each other. This confirmed the suggestion [39], [40] that the results of the UMUX-LITE can predict those of the SUS in scenarios with a low number of questions. On the other hand, the dimension of *pragmatic* quality of the UEQ-S also showed high to very high correlations with

the SUS and UMUX-LITE. However, there were differences in the correlations with the *hedonic* dimension, which was to be expected as usability questionnaires are aimed at *pragmatic* quality.

In 2020 [41], the correlations between five UX factors of the UEQ+, the complete SUS and the Net Promoter Score (NPS) were also examined. The NPS is used to measure customer loyalty and is not a classic usability questionnaire, but suggests that poorly usable products would not maintain a high NPS. The factors *Intuitive Use*, *Quality of content*, *Trustworthiness of content*, *Trust* and *Stimulation* were selected as the equivalent of the UEQ+ for the comparison of the evaluation of a product. The key performance indicators of the questionnaires were calculated on the basis of the results. It was found that the NPS correlates positively and linearly with the SUS and UEQ+ KPI moderator. The SUS and UEQ+ KPI also correlate strongly with each other in a positive linear fashion, especially the UX factor *Intuitive use*.

In summary, it can therefore be said that similarities between the *pragmatic* factors of the UEQ, UEQ-S and UEQ+ and classic usability questionnaires have already been identified in the past. The UEQ family can therefore be used as a replacement or supplement to comparable established usability questionnaires (e.g., SUS, UMUX, ISONORM). In addition, based on a more holistic model of user experience, it measures UX factors that cannot be measured completely and/or in several languages by other UX questionnaires (e.g., AttrakDiff2, VISAWI, meCUE).

IV. INTERPRETING RESULTS FROM UX QUESTIONNAIRES

How is it that the same product is sometimes perceived very differently by separate individuals? On the one hand, general influences such as demographic factors (e.g., age or gender) or cultural background matter. The different levels of experience already gained with the product can also have an influence.

On the other hand, the significance of the measured values is tangible if physical properties of objects are to be measured, such as the weight of an object or the reaction time of a system to an input. However, the situation is different with psychological properties, such as aesthetics or stimulation. These concepts need to be measured just as carefully in order to make different individual opinions tangible.

For this reason, the recommended statistical instruments of the UEQ Family are presented below, which are necessary to make the different results both comparable and, above all, interpretable.

A. Statistical Analysis

Once the subjective UX has been recorded using the questionnaires, it is necessary to evaluate and interpret the results. To make it easier to get started, Excel sheets are provided for the UEQ family, depending on the questionnaire version used (UEQ [30], UEQ-S [42], UEQ+ [37]). The raw data, i.e. the cell-by-cell scores between -3 and +3, are inserted into this sheet. The tool then first returns the mean values per participant, item and factor, with the corresponding position (standard deviation) and spread (variance). Based on this, also the confidence interval is calculated in the tool. The factor means and confidence intervals are shown as an example in Fig. 9.

It is clear that every factor within the range from -3 to 3 was rated higher than 0.7. This means that the overall UX of the product was perceived as rather positive (green area). Nevertheless, the *pragmatic* factors *Perspicuity* and *Efficiency* were rated best for this fictitious product, while the *hedonic* factors *Stimulation* and *Novelty* were rated worst and are still partially in the neutral range (yellow). On the one hand, these are indications on which product features require improvement. On the other hand, the use case of the product should be considered at this point. Products with a focus on achieving *pragmatic*

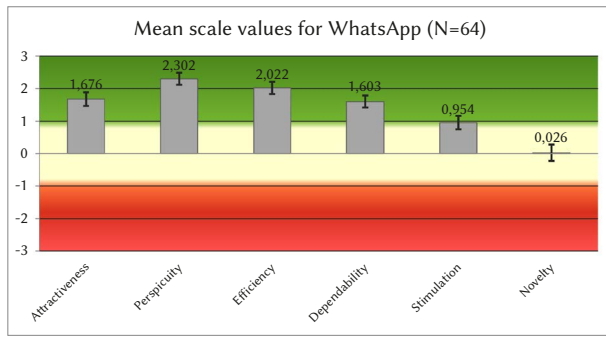


Fig. 9. Product evaluation for WhatsApp (N=64) [25], measured with the UEQ and evaluated with the UEQ Excel tool [30].

goals (e.g., word processing with Microsoft Word) should, as expected, score better on the *pragmatic* factors, while leisure products with a focus on entertainment (e.g., social networks such as TikTok) should expectedly score better on *hedonic* factors. The confidence intervals (lines at the edge of the bars) in Fig. 9 are small (< 0.25). Since not all actual users can be reached in the UX measurement, but the surveys are carried out with the help of random samples, this assessment is necessary in order to draw conclusions about the range in which the actual factor mean lies. The smaller the line, the more likely a similar result is for other samples. If, on the other hand, the lines of the confidence interval are wide, the results should not be overinterpreted. This can occur in the case of small sample sizes or several outliers (e.g. a few very good or very poor ratings by individual participants), making further investigations necessary. In this case, for example, the examination of inconsistencies in the Excel tools can be used, which provides information on whether some answers may not have been answered seriously by participants. This is the case if several items on the same factor were answered very differently (more than 3 points different) (see Fig. 10).

Scales with inconsistent answers		
Pragmatic Quality	Hedonic Quality	Critical?
1		1
		0
1		1
		0
		0
		0
		0
1	1	2

Critical length	
Same answer for	Middle Category
2	
4	
3	
3	
5	
8	Remove
8	
3	

Fig. 10. Example of inconsistencies in answers per factor, measured with the UEQ-S and calculated with the UEQ-S Excel tool [42].

In this fictitious excerpt of responses, it is clear that inconsistencies were found in the answers of three participants. One participant also answered inconsistently in items of both dimensions, which should be critically examined. In addition, two participants ticked the same answer option for all items. One of these two participants only selected the middle option (neutral, 0), which is why it is also recommended to remove answers of this participant.

Another way of assessing how accurate the UX measurements are is reliability, measured using the established statistical method of Cronbach's alpha. The calculations for this are also provided in the Excel tools of the UEQ Family [30], [37], [42]. An example of this is shown in Fig. 11.

The individual values in the Cronbach's alpha column indicate how reliably the questionnaire (in Fig. 11: the UEQ+) measures the respective factors with the items. It can be stated that values > 0.7 are acceptable, while values > 0.8 are good and > 0.9 are very good. As all alpha values in this fictitious example are greater than 0.87, it can be assumed that the measurement accuracy is high. The remaining

columns show the correlation of the individual factors. Correlation values of less than 0.9 are desirable here. Values above 0.9 would indicate that individual questions may be too similar to each other or not sufficiently selective.

Scale	Corr(11,12)	Corr(11,13)	Corr(12,13)	Cronbach Alpha
Attractiveness	0,79	0,63	0,66	0,87
Efficiency	0,75	0,81	0,88	0,91
Intuitive Use	0,79	0,85	0,93	0,94
Visual Aesthetics	0,76	0,85	0,85	0,95
Quality of Content	0,46	0,43	0,63	0,83
Trustworthiness of Content	0,58	0,57	0,89	0,90
Social interaction	0,77	0,76	0,87	0,94

Fig. 11. Example of a reliability in answers per factor, measured with the UEQ+ and calculated with the UEQ+ Excel tool [37].

The statistical analyses of the Excel tool can therefore be used to make initial interpretations of the performance of the factors and items, and therefore features of the product under investigation. Nevertheless, in practice, decision-makers are used to having the important information about the company's products and situations summarized in an overall key figure, rather than having to perform various statistical calculations. In addition, as the UEQ+ research basis (see Section III.C) has already shown, different factors are relevant for each product. The question therefore arises as to how the overall UX performance can be summarized, how the importance can be measured and then set in relation to this performance.

B. Benchmark

In addition to the actual statistical measurements relating to how users perceive the UX of products, the first fundamental question in practice is also whether the expectations set in advance have been met. Assuming that the previous statistical measurements produce values and error bars as shown in Fig. 9, the question arises as to whether this is a good or bad result.

It is easy to compare a product version with a previous version to compare expectations and measurements. The UEQ family also provides a separate tool for this purpose free of charge online [30]. However, it becomes more difficult if no previous version is available for comparison, but the question arises as to whether the product has a sufficient UX. It is therefore interesting to compare the extent to which the measured UX of a product matches the measured UX of other products, measured using the same method. Some questionnaires provide benchmarks for this purpose, which contain the UX scores measured with the questionnaire based on a large number of different products. Benchmarks were developed for the UEQ Family in order to provide a greater basis for interpreting the product-specific results.

1. UEQ Benchmark

A benchmark for the UEQ was developed by Schrepp et al. in 2014 and updated in 2017 [3]. Data from a total of 246 products and 9,905 responses were used as a basis for evaluation, including for example business applications, web stores or services, social networks, household appliances and other product types. Due to the high degree of confidence in the actual data, the factor averages were integrated into the data set instead of raw data.

Based on this data set, the UEQ benchmark provides a grouping per factor in the following 5 categories, meaning the evaluated product is

- **Excellent:** among the best 10% of the product base.
- **Good:** 10% of products are better and 75% are worse.
- **Above Average:** 25% of products are better and 50% are worse.
- **Below Average:** 50% of products are better and 25% are worse.
- **Bad:** among the worst 25% of the product base.

With regard to the benchmark dataset basis, 20-30 users already produce stable results. However, general UX expectations change over time. So even if the underlying data set is not continuously updated, new products can still achieve *good* scores [3].

In relation to the results of Fig. 9, this leads to a benchmark corresponding to Fig. 12. The benchmark can be calculated using the UEQ Excel sheet [30].

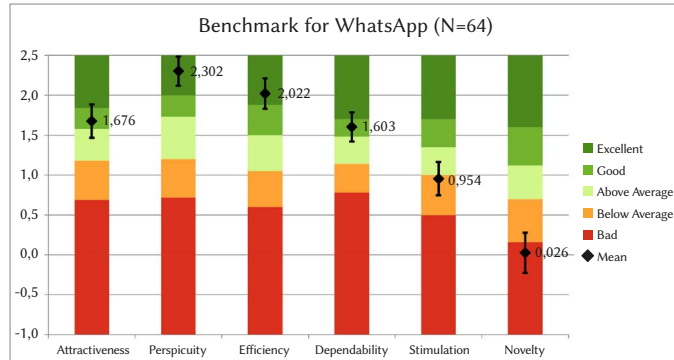


Fig. 12. Product benchmark for WhatsApp (N=64) [25], measured with the UEQ and evaluated with the Excel tool [30].

Compared to the other products of the benchmark dataset, it is clear that the *pragmatic* factors perform in the categories *good* or *excellent*, while the *hedonic* factors can be classified as *below average* and even *bad*. This may be related to the intended use of the product. Even if WhatsApp can be used in a leisure environment, it serves the fulfillment of purposes or completing tasks, such as efficient communication, so that the focus of the product is on the pragmatic properties. For this reason, impact factors (see Section V) should also be taken into account when interpreting the benchmark.

2. UEQ-S Benchmark

As much of the data for the UEQ benchmark comes from practical industrial projects, for confidential reasons the factor means rather than the raw data were included in the creation of the UEQ benchmark. As analyzed by Hinderks et al. [43], it is therefore not possible to synthesize a benchmark exclusively for the 8 items of the UEQ-S on the basis of the data set. However, as it was shown (see Section III.B) that the UEQ values allow a good approximation of the UEQ-S values, the question arose as to whether the UEQ-S benchmark can be calculated on the basis of the UEQ benchmark. For this purpose, the mean values of the factors *Efficiency*, *Perspicuity* and *Dependability* were calculated in a study in 2018 to form the *pragmatic* quality dimension, and the mean values of the *Stimulation* and *Novelty* factors were calculated for the *hedonic* dimension. The overall value was formed on the basis of all UEQ factors, including *Attractiveness*. The following results were thus obtained for the dimensions of the *pragmatic* and *hedonic* quality (PQ and HQ) of the UEQ-S benchmark:

- **Excellent:** PQ greater than 1.73, HQ greater than 1.55, Overall greater than 1.58.
- **Good:** PQ between 1.55 and 1.73, HQ between 1.25 and 1.55, Overall between 1.4 and 1.58.
- **Above Average:** PQ between 1.15 and 1.54, HQ between 0.88 and 1.24, Overall between 1.02 and 1.39.
- **Below Average:** PQ between 0.73 and 1.14, HQ between 0.57 and 0.87, Overall between 0.68 and 1.01.
- **Bad:** PQ less than 0.73, HQ less than 0.57, Overall less than 0.68.

The benchmark can be calculated again using the corresponding UEQ-S Excel sheet [42]. The results showed that a natural

transformation of the UEQ to the UEQ-S is possible in this way. However, over-interpretation is not advisable, as only the factor means were used as a basis for the analyses. More data would be needed to gain a deeper understanding of the relationship between full and short UEQ version.

3. UEQ+ Benchmark

Due to the modular structure of the UEQ+, it is difficult and time-consuming to create a classic benchmark at the level of the individual factors, as only some factors are relevant for certain products. It would require significantly more products and data than when creating the UEQ benchmark. In addition, further UEQ+ factors can be added over time, meaning that a UEQ+ benchmark can never assess all factors with the same quality.

A study from 2023 by Meiners et al. [44] therefore took a different approach and carried out a simple benchmark based on a limited set of product evaluations as quick guidance for UX researchers and practitioners. This benchmark is based on the UX KPI and therefore not on individual factors, but on the overall UX impression of a product. These KPIs also vary depending on the selected factors and products, but the comparison of several products with the same use case is often sufficient in practice to get a first impression of whether one's own product is "good enough" in terms of its UX. Accordingly, for the calculation of the UEQ+ benchmark, which is based on the UEQ+ KPI, various products were evaluated by over 3,200 participants in a total of 26 studies in order to allow an initial understanding of the perceived compared to the expected UX quality of products. Further information can be found in Meiners et al. [44]. The required KPIs can be calculated using the UEQ+ Excel sheet [37]. However, this first UEQ+ benchmark should not be overinterpreted, as this would require more data.

C. Key Performance Indicators

To further effectively evaluate the results of an UX questionnaire, there is a relevant desire for a meaningful key figure, a so-called key performance indicator, as an assessment of the perceived UX of a product. A KPI helps in identifying important areas for improvement, and the multidimensional nature of the user experience itself already returns multiple scales that provide information about these areas [26]. To now further reduce the complexity of multidimensionality, especially for decision-makers, a UX KPI for the UEQ Family is introduced. In order to enable an assessment of strengths and weaknesses, so-called importance ratings were added to the UEQ from 2019 [26], which respondents use to assess the relevance of the respective factors and associated items for a product. These importance ratings are recorded in a range from -3 to 3. Based on these ratings, the relative importance of each factor can then be calculated for each participant. In order to use a key figure to estimate how strong the correlation is between the perceived quality (the value measured with the UEQ factor) and the perceived importance for the respondents, a dependency is initially assumed. This is understandable, because the more important an item is for a user, the more seriously they will answer it in relation to the product. This can also be seen in Fig. 13.

The *pragmatic* factors of the WhatsApp were rated as more important than the *hedonic* factors. The mean values of the *pragmatic* factors were also rated better than those of the *hedonic* factors. In order to obtain a meaningful overall impression of the product taking these considerations into account, the relative rating of each UEQ factor (measured with the UEQ questions) is multiplied by the relative importance of the same factor (measured with the importance questions), and the individual results are added up and divided by the number of participants. The full explanation can be found in the corresponding study [26].

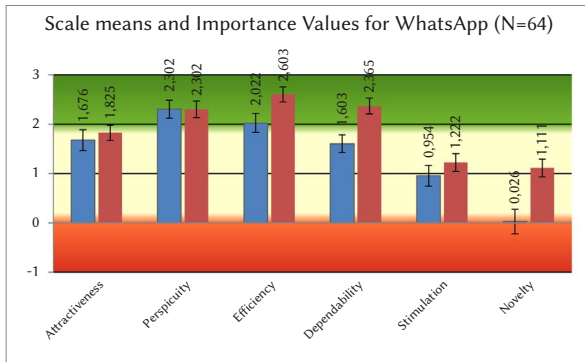


Fig. 13. Product evaluation for WhatsApp (N=64) with factor means (blue) and importance ratings (brown) [25], measured with the UEQ and evaluated with the UEQ Excel tool [30].

The result is again a value in the range from -3 to 3. To simplify the calculation of the KPI, this calculation was also integrated into the Excel tools [30], [37], [42] of the UEQ Family.

However, if we now consider that we want to have a meaningful KPI as an evaluation of the overall UX, then a number alone is not meaningful enough. For example, how do the KPI values 1.1 and 1.4 differ? This raises the question of how an interpretation guideline can be provided for practitioners and researchers. Calculated on the basis of the benchmarks (for further information see [45]), the KPI value range according to Table I can be determined.

TABLE I. KPI VALUE RANGE (SCALE FROM -3 TO 3)

UEQ Benchmark	UEQ KPI Min Value	UEQ KPI Max Value	UEQ KPI Mean Value
Excellent	2.018	2.143	2.080
Good	1.375	1.628	1.502
Above Average	1.038	1.290	1.164
Below Average	0.645	0.891	0.768
Bad	-0.286	-0.162	-0.224

It can therefore be summarized that although values from -3 to +3 can theoretically occur, this will most likely not happen in the practical application of the UEQ KPI. The actual range extends from -0.286 to +2.080 and is therefore smaller and more positive. On the basis of this guide, a first impression of the UX of the product or use case under consideration can be gained. However, this key figure should also be treated with caution, as there can be major differences in the importance and evaluation of the actual factors, the performance of the product. For example, a lower rating of the *pragmatic* factors compared to *hedonic* factors, but with a reversed distribution of importance, could also indicate specific weaknesses in the product [26], [45]. It is therefore advisable to take a closer look at the results as part of further analyses.

D. Importance-Performance Analysis

The importance-performance analysis (IPA) [46], [47] is a graphical representation of the relationship between relevance and actual UX assessment of individual factors by participants. The aim of this analysis is to identify specific recommendations for action for the individual factors. It is assumed that a user is satisfied if the perceived importance (brown bar in Fig. 13) has been fulfilled. The fulfillment (performance) is expressed by the factor mean value (blue bar in Fig. 13). This means that it is not the absolute difference between importance and performance which is relevant, but the relative difference between them.

There is no prescribed list of factors for this, which is why this analysis can be carried out with all factors of the UEQ family

questionnaires, provided that the importance ratings of the individual factors are also recorded for these factors in the form of a 7-point Likert rating.

According to Hinderks et al. [47], the mean values for each factor are presented in an IPA plot with a total of four quadrants, with each factor being assigned a point. This point is calculated by using the performance value (factor mean value of the perceived UX) for the x-axis and the importance value (mean value of the importance questions) for the y-axis. In this way, the individual quadrants represent concrete recommendations for action, which are described in Table II.

TABLE II. IPA QUADRANTS [46]–[48]

IPA quadrant	description
Keep Up the Good Work	great strengths and potential competitive advantages both importance and performance equally highly rated no need for action
Possible Overkill	factors rated relatively low, importance is below performance further development of factors not necessary / inefficient
Low Priority	low importance and performance, no action required, balanced
Concentrate Here	most important relatively important, while performance below average [47] highest potential for improvement

An example implementation of this IPA is shown in Fig. 14. Each point in this plot represents a selected factor whose importance and performance (factor mean) was measured with the UEQ+. While the dashed axes represent the original coordinate origin, the solid axes represent the quadrants that are necessary for the IPA interpretation. These are formed by the mean value of all the factors shown and considered [1], [46], [47].

If the importance is higher than the performance, this factor should be improved. In Fig. 14, this applies in particular to the *hedonic* factor *Novelty*. However, no changes are necessary for the *Efficiency* and *Attractiveness* factors. Nevertheless, external factors must also be taken into account when interpreting the results, as can be seen in Fig. 15.

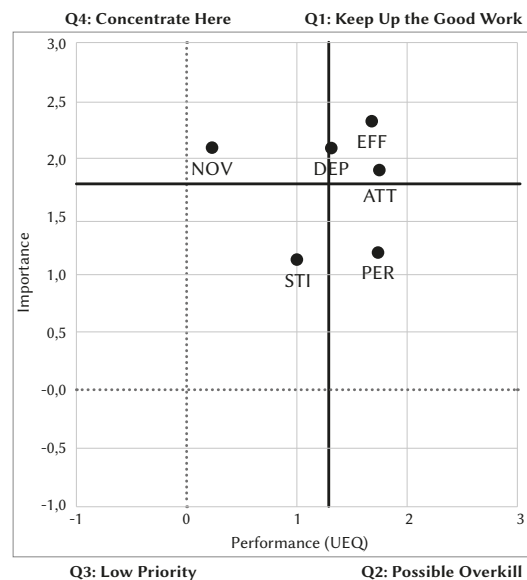


Fig. 14. IPA plot for the product WhatsApp and the factors *Novelty* (NOV), *Stimulation* (STI), *Dependability* (DEP), *Efficiency* (EFF), *Attractiveness* (ATT) and *Perspicuity* (PER), measured with the UEQ+ [1], [46]–[48].

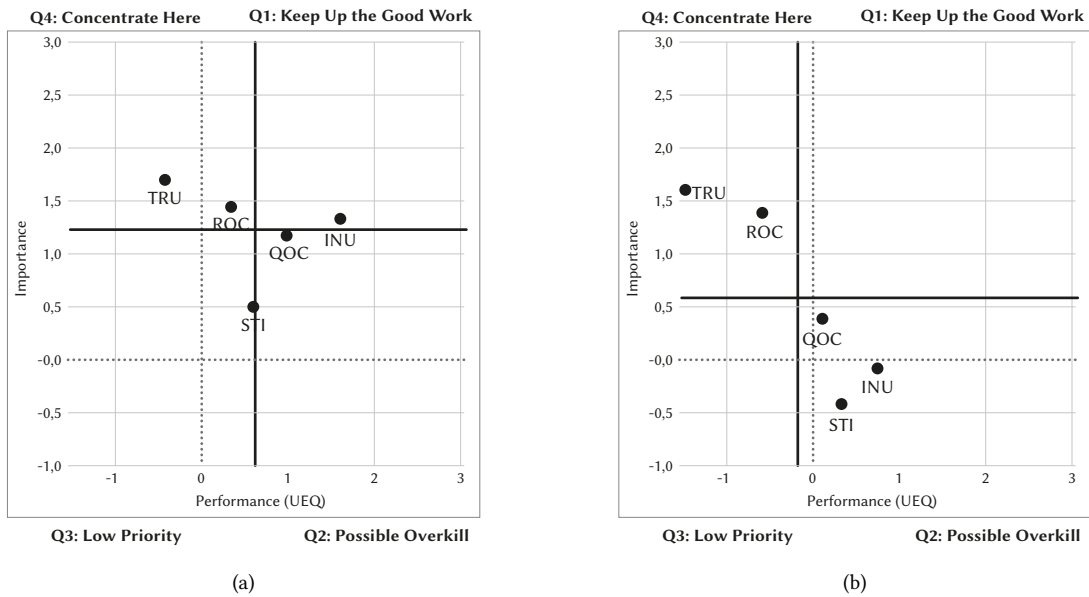


Fig. 15. IPA plot for the product Facebook and the factors *Novelty* (NOV), *Stimulation* (STI), *Dependability* (DEP), *Efficiency* (EFF), *Attractiveness* (ATT) and *Perspicuity* (PER), measured with the UEQ+ [1], [46]–[48].

Fig. 15 compares two IPAs, where variant (a) was created on the basis of users who use the fictitious product every day and variant (b) on the basis of users who do not use it every day. It is clear that the frequency of use can therefore influence the interpretation of the statistical analyses, benchmark, KPI and IPA, and thus the recommendations for action. For this reason, relevant impact factors that are necessary for understanding the UX measurement with the UEQ family are presented in the following Section.

V. IMPACT FACTORS

Over the years, various studies have been carried out on impact factors on the perceived user experience. A number of aspects were found to have an influence.

A. User-Related Influences

Starting with the factor **gender**, it cannot be ruled out that, depending on the product type, this may have an influence on the perceived user experience measured with the UEQ, even if it can be small [49], [50]. These results were also confirmed in comparison with other questionnaires. Therefore, if a specific target group is defined on the basis of gender, it should be ensured in advance that the expectations can be met accordingly.

The situation is similar with external influencing factors such as **duration of use**, **frequency of use** and **knowledge** of the product [50], [51]. It can be assumed that users who use their product frequently typically know it better and adapt their usage behavior accordingly in order to avoid typical UX problems with the product. Accordingly, the perceived user experience can also change, which could be confirmed on the basis of various products. The influences of frequency of use and knowledge have been proven, but these also vary depending on the specific product. Accordingly, it is also advisable to select the right factors for each product or product category in advance in order to make the user experience more suitable (see Section V).

The **cultural influence** should not be ignored either. As various studies have shown, there are differences in the perceived UX, which can be explained by different cultures. This should be distinguished from the perceived importance of UX factors, which were shown in these studies to be dependent on the product or product category

(see Section V), as there were clear similarities between the cultures investigated. However, differences in perception within the German vs. the Indonesian culture were found, even if they were only minor. The greatest differences were found in the perception of *hedonic* factors such as *Identity*, *Novelty* and *Stimulation*. This was demonstrated with the help of the proportion of variance, which was proven to be generally higher for *hedonic* than *pragmatic* UX factors. These *hedonic* factors were rated relatively low by the German participants compared to the *pragmatic* factors. Nevertheless, also the extent of the influence of culture compared to other interindividual differences between persons was examined here. It was found that the influence of culture is relatively low compared to the impact of differences on an individual level. On the one hand, this means that culture is an impact factor that should not be ignored when measuring UX. On the other hand, however, it should not be overestimated, as personal preferences predominate. Studies on the influence of other cultures should therefore be carried out before further conclusions are drawn here [52]–[54].

Finally, there are influences that relate to the structure of the questionnaires depending on the person who is filling it out. The circumstances are different in contrast to the personal influences, for example with regard to the **item polarity** of semantic differentials. While the UEQ has a mixed item polarity, in which half of the positive items (e.g. entertaining, easy to use) are randomized to the right and the other half to the left, the UEQ-S and UEQ+ have a one-sided (right-sided) item polarity. The basic idea of the study by Schrepp et al. in 2023 [25] was to simplify the processing of the UEQ for the participants by using a one-sided variant. However, studies have shown, also in comparison with other questionnaires, that a change in polarity of the UEQ to a one-sided variant has more disadvantages than advantages. Even if the number of inconsistencies, clicks and the processing time is slightly lower with a one-sided variant, these effects were proven to be very small practically irrelevant [25]. Furthermore, a response tendency is introduced in which participants tend to place items on the same side if they are uncertain about individual items (e.g., on the right if they have a positive perception of the product). Due to this conditional influence of item polarity, a change in the UEQ should therefore be avoided [25].

B. Product-Related Influences

In addition to the influencing factors that relate to the users and their environment, influencing factors that relate to the products under consideration are also clearly present. It has been proven that the user experience is evaluated differently depending on the **product** or **product category**. UX is generally difficult to measure not only because different users subjectively evaluate the same product differently, but also because the same user evaluates different products differently. This is due to the association of products with certain characteristics. For example, while the *Quality of Content* is not applicable for word processing tools, it is most relevant for news portals. The factor of *Stimulation* is also rated as less relevant for online banking platforms, in contrast to social networks (see Fig. 16). This must be taken into account both when measuring and evaluating the perceived UX using the UEQ Family.

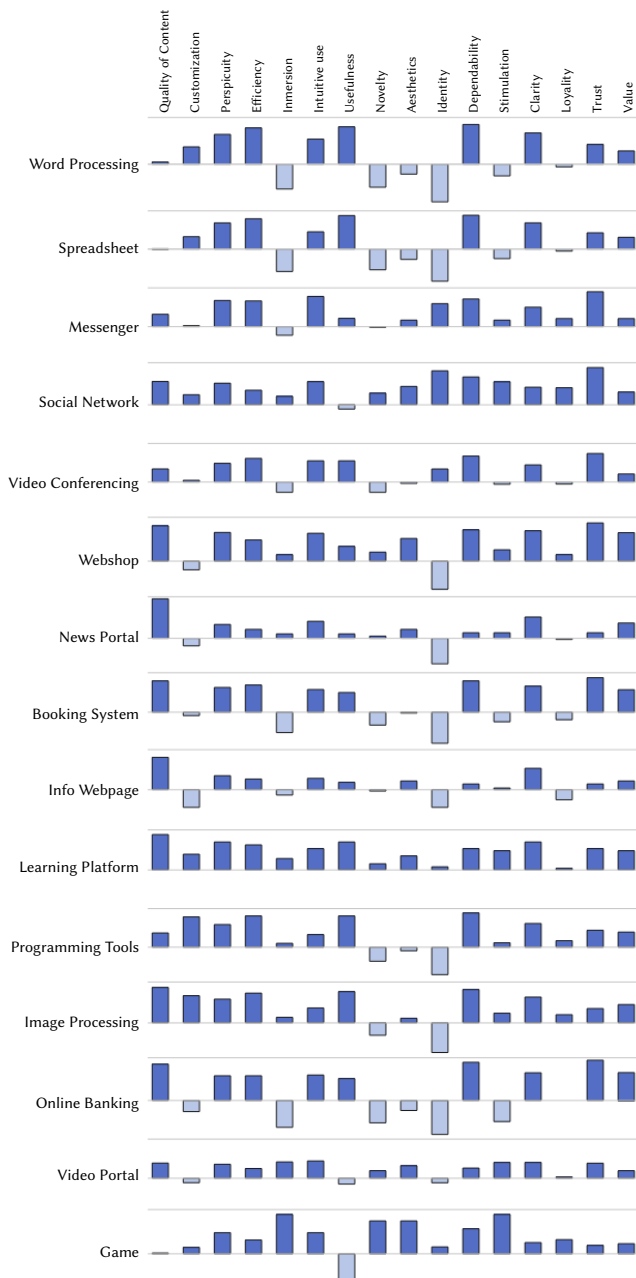


Fig. 16. Means of the importance ratings for the UX quality aspects per product category [32]. Scale ranges from -3 to +3.

These importance ratings can also be transferred to specific products. Studies have shown high to almost perfect correlations between the importance ratings of specific products and their associated product categories. These include, for example, Microsoft Word in the Text Processing category (see Fig. 17), WhatsApp in the Messengers category, or YouTube and Netflix in the Video Portals category [55].

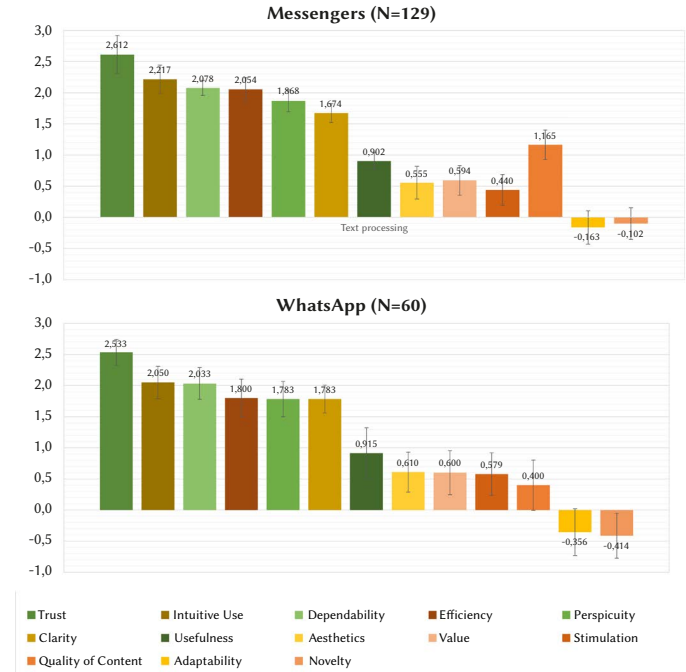


Fig. 17. Importance ratings for the product WhatsApp and the product type Word Messengers. Scale ranges from -3 to +3.

A similar analysis was carried out for the product category of collaboration tools. Here, for example, the UX factors *Trust*, *Perspicuity* and *Efficiency* were rated as the most important overall. The individual differentiations of the products themselves, e.g., Microsoft Teams or Discord, did not impact the rating of the UX aspects as much [56].

The question that now arises is how these impact factors and interpretation notes can be taken into account in practice.

VI. CONCLUSION AND FUTURE WORK

Measuring user experience is useful in order to record ongoing improvements to products, services and systems that may result from continuous bug fixes or new releases. Furthermore, it is essential for the product success to determine whether the UX is sufficient, where there are areas for improvement and also whether the investments in the UX are worthwhile in terms of return on investment [57]. UX heuristics already form a suitable basis and guideline for this [18].

Nevertheless, it can be difficult to decide what exactly should be measured. On the one hand, users have different subjective and product-dependent opinions. On the other hand, it is necessary to determine what exactly is meant by UX and to select the appropriate factors for the use case and the research questions from a large number of possible factors. For this reason, standardized questionnaires such as the UEQ, UEQ-S and UEQ+ help to determine what exactly a user thinks about a product and how the generally perceived UX can be assessed.

With the help of semantic differentials, the UEQ forms a good basis for a broad range of use cases and at the same time provides

meaningful possibilities for interpretation. The UEQ is currently available in more than 30 languages [33] and also in a simplified language form, e.g., for children [28].

However, it is still possible that the UEQ takes up too much time for the specific target group. In this case, the UEQ-S can save time, but provides a less detailed picture. In cases where a comprehensive impression for a more specific research question is needed, the usage of the UEQ+ is recommended. Modular questionnaires like the UEQ+ are suitable if either a complete picture of a product needs to be formed, or also only specific UX factors need to be considered. With the help of the UEQ+, it is possible to avoid using other questionnaires in addition to the UEQ in order to capture more than the predefined factors. It also makes it possible to select specific factors depending on the product. Also the UEQ+ is currently available in more than 20 languages, so that it offers potential for versatile use.

However, both when using modular and holistic questionnaires from the UEQ Family, is recommended to consider in advance which product or product category is to be evaluated or given priority. Existing studies provide indications of the importance of certain UX factors for specific products and categories. These are advisable to be used when selecting the appropriate factors. Even if individual products have possibly not yet been considered in explicit studies, it has been confirmed that the UX factors relevant to certain product categories can be used to draw conclusions about individual products [32], [55] and are among the good practices of creating a common UX vision and shared understanding [58]. These findings can also be integrated into internal company processes. For example, the selection of suitable UX factors for each product can be integrated into agile methods such as UX Poker [59]. This method takes place early on in the process and facilitates a shared understanding of UX within the team by assessing the influence of user stories based on selected UX factors. The results of the UX evaluation can thus contribute directly to fulfilling the requirements of UX management to create a positive UX [60].

Once the appropriate factors have been selected and the user data collected, it is necessary to statistically analyze the results. Section IV and the Excel tools provided by the UEQ family [30], [37], [42] offer guidance on how to interpret the results appropriately. This means that no previous experience with questionnaires is necessary to facilitate the use and implementation of the UEQ Family. In addition, however, this article presents common interpretation methods (e.g. KPI and IPA) as well as concrete interpretation aids (e.g. benchmark) for the results.

In addition, the UEQ family can both supplement questionnaires already in use (e.g. on usability aspects or KPIs) and replace them (e.g. by providing a more holistic picture through the addition of a hedonic perspective). Existing results often do not have to be discarded, but can be derived using the UEQ (e.g. SUS or UMUX-LITE, see Section III.D). The UEQ family can therefore be used in a variety of ways depending on the research question and company context.

REFERENCES

- [1] A. Hinderks, D. Winter, M. Schrepp, J. Thomaschewski, "Applicability of user experience and usability questionnaires," *Journal of Universal Computer Science*, no. 25, pp. 1717–1735, 2020.
- [2] E. L.-C. Law, P. van Schaik, "To measure or not to measure ux: An interview study," in *International Workshop on the Interplay between User Experience Evaluation and System Development (I-UxSED)*, 2012.
- [3] M. Schrepp, J. Thomaschewski, A. Hinderks, "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.
- [4] M. Schrepp, A. Hinderks, J. Thomaschewski, "Applying the user experience questionnaire (UEQ) in different evaluation scenarios," in *Design, User Experience, and Usability: Theories, Methods, and Tools for Designing the User Experience*, Cham, 2014, pp. 383–392, Springer International Publishing. isbn: 978-3-319-07668-3.
- [5] M. Schrepp, J. Thomaschewski, "UEQ+ Handbook." https://ueqplus.ueq-research.org/Material/UEQ+_Handbook_V6.pdf, 2023. [Online; accessed 24-April-2024].
- [6] M. Schrepp, *User Experience Questionnaires: How to use questionnaires to measure the user experience of your products?* Self-published, 2021. isbn: 979-8736459766.
- [7] M. Schrepp, A. Hinderks, J. Thomaschewski, "Design and evaluation of a short version of the user experience questionnaire (UEQ-s)," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 6, p. 103, 2017, doi: 10.9781/ijimai.2017.09.001.
- [8] B. Laugwitz, T. Held, M. Schrepp, "Construction and evaluation of a user experience questionnaire," in *HCI and Usability for Education and Work*, vol. 5298 of *Lecture Notes in Computer Science*, A. Holzinger Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 63–76, doi: 10.1007/978-3-540-89350-9_6.
- [9] M. Schrepp, J. Thomaschewski, "Design and validation of a framework for the creation of user experience questionnaires," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 7, 2019, doi: 10.9781/ijimai.2019.06.006.
- [10] DIN EN ISO 9241-11, "Ergonomics of human-system interaction — part 11: Usability: Definitions and concepts," 2018. doi: 10.31030/2757945.
- [11] ISO9241-210, "Ergonomics of human-system interaction - part 210: Human-centred design for interactive systems," 2020. doi: 10.31030/3104744.
- [12] M. Hassenzahl, "The thing and i: Understanding the relationship between user and product," in *Funology*, M. A. Blythe Ed., Boston [etc.]: Kluwer Academic Publishers, 2003, pp. 31–42. isbn: 978-1402029660.
- [13] M. Schrepp, J. Kollmorgen, J. Thomaschewski, "A comparison of sus, umux-lite, and UEQ-s," *Journal of User Experience*, vol. 18, p. 86–104, jun 2023.
- [14] M. P. Cota, J. Thomaschewski, M. Schrepp, R. Gonçalves, "Efficient measurement of the user experience. a portuguese version," *Procedia Computer Science*, vol. 27, pp. 491–498, 2014, doi: 10.1016/j.procs.2014.02.053.
- [15] V. Roto, E. L.-C. Law, A. Vermeeren, J. Hoonhout, "User experience white paper: Bringing clarity to the concept of user experience," Schloss Dagstuhl - Leibniz- Zentrum für Informatik, 2011.
- [16] E. M. Schön, J. Hellmers, J. Thomaschewski, "Usability evaluation methods for special interest internet information services," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 2, no. 6, pp. 26–32, 2014, doi: 10.9781/ijimai.2014.263.
- [17] J. Nielsen, R. Molich, "Heuristic evaluation of user interfaces," in *CHI '90: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1990, pp. 249–256. isbn: 0201509326.
- [18] F. Bader, E.-M. Schön, J. Thomaschewski, "Heuristics considering ux and quality criteria for heuristics," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 6, pp. 48–53, 2017, doi: 10.9781/ijimai.2017.05.001.
- [19] M. Rauschenberger, M. Schrepp, M. Perez- Cota, S. Olschner, 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, p. 39, 2013, doi: 10.9781/ijimai.2013.215.
- [20] M. Schrepp, "A comparison of ux questionnaires - what is their underlying concept of user experience?," *Mensch und Computer 2020 - Workshopband*, Bonn, 2020. doi: 10.18420/muc2020-ws105-236.
- [21] J. Brooke, "Sus: A quick and dirty usability scale," in *Usability Evaluation In Industry*, CRC Press, 1996, p. 6. isbn: 9780429157011.
- [22] M. Schrepp, "UEQ Handbook." <https://www.ueq-online.org/Material/Handbook.pdf>, 2023. [Online; accessed 24-April-2024].
- [23] M. Hassenzahl, M. Burmester, F. Koller, "Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität," in *Mensch & Computer 2003*, Stuttgart [u.a.], 2003, pp. 187–196, Teubner. isbn: 978-3519004417.
- [24] M. Minge, L. Riedel, "mecue - ein modularer fragebogen zur erfassung des nutzungserlebens," in *Mensch & Computer 2013: Interaktive*

- Vielfalt, München, 2013, pp. 89–98, Oldenbourg Verlag. doi: 10.1524/9783486781229.89.
- [25] M. Schrepp, J. Kollmorgen, J. Thomaschewski, "Impact of item polarity on the scales of the user experience questionnaire (ueq)," in *International Conference on Web Information Systems and Technologies*, 2023, pp. 15–25. doi: 10.5220/0012159900003584.
- [26] A. Hinderks, M. Schrepp, F. J. Domínguez Mayo, M. J. Escalona, J. Thomaschewski, "Developing a ux kpi based on the user experience questionnaire," *Computer Standards & Interfaces*, vol. 65, pp. 38–44, 2019, doi: <https://doi.org/10.1016/j.csi.2019.01.007>.
- [27] A. Hinderks, M. Schrepp, J. Thomaschewski, "UEQ and UEQ-S Website." <https://www.ueq-online.org/>, 2024. [Online; accessed 24-April-2024].
- [28] M. Schrepp, M. P. Cota, R. M. Gonçalves, A. Hinderks, J. Thomaschewski, "Adaption of user experience questionnaires for different user groups," *Universal Access in the Information Society*, vol. 16, pp. 629 – 640, 2016, doi: 10.1007/s10209-016-0485-9.
- [29] M. Hernández-Campos, J. Thomaschewski, Y. C. Law, "Results of a study to improve the spanish version of the user experience questionnaire (UEQ)," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4, pp. 202–207, 2023, doi: 10.9781/ijimai.2022.11.003.
- [30] A. Hinderks, M. Schrepp, J. Thomaschewski, "UEQ Data Analysis Tool." https://www.ueq-online.org/Material/Data_Analysis_Tools.zip, 2024. [Online; accessed 24-April-2024].
- [31] A. Hinderks, M. Schrepp, M. Rauschenberger, J. Thomaschewski, "Reconstruction and validation of the ux factor trust for the user experience questionnaire plus (ueq+)," in *Proceedings of the 19th International Conference on Web Information Systems and Technologies - Volume 1: WEBIST*, 2023, pp. 319–329, INSTICC, SciTePress. doi: 10.5220/0012186700003584.
- [32] M. Schrepp, J. Kollmorgen, A. Meiners, A. Hinderks, D. Winter, H. Santoso, J. Thomaschewski, "On the importance of ux quality aspects for different product categories," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 2, pp. 232–246, 2023, doi: 10.9781/ijimai.2023.03.001.
- [33] M. Schrepp, J. Thomaschewski, "UEQ+ Website." <https://ueqplus.ueq-research.org/>, 2024. [Online; accessed 24-April-2024].
- [34] A. M. Klein, A. Hinderks, M. Schrepp, J. Thomaschewski, "Construction of ueq+ scales for voice quality: measuring user experience quality of voice interaction," in *Mensch und Computer 2020 - Tagungsband*, New York: ACM, 2020, p. 1–5, doi: 10.1145/3404983.3410003.
- [35] M. Lindemann, H. Weber, "Ux in zahlen: Der user experience questionnaire for medical devices," in *Usability Professionals 23*, Gesellschaft für Informatik e.V., 2023, doi: 10.18420/muc2023-up-455.
- [36] E. Mortazavi, P. Doyon-Poulin, D. Imbeau, J.-M. Robert, "Development and validation of four social scales for the ux evaluation of interactive products," *International Journal of Human-Computer Interaction*, pp. 1–14, 2023, doi: 10.1080/10447318.2023.2258026.
- [37] M. Schrepp, J. Thomaschewski, "UEQ+ Data Analysis Tool." https://ueqplus.ueq-research.org/Material/UEQ_Plus_Data_Analysis_Tool.xlsx, 2024. [Online; accessed 24-April-2024].
- [38] A. Hinderks, M. Schrepp, J. Thomaschewski, "Vergleich von ux fragebögen." doi: 10.18420/muc2018-mci-0363.
- [39] J. R. Lewis, B. S. Utesch, D. E. Maher, "Umux-lite: when there's no time for the sus," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, New York, NY, USA, 2013, p. 2099–2102, Association for Computing Machinery. doi: 10.1145/2470654.2481287.
- [40] U. Lah, J. R. Lewis, B. Sumak, "Perceived usability and the modified technology acceptance model," *International Journal of Human-Computer Interaction*, vol. 36, pp. 1216 – 1230, 2020, doi: 10.1080/10447318.2020.1727262.
- [41] A.-L. Meiners, A. Hinderks, J. Thomaschewski, "Korrelationen zwischen ux-fragebögen." *Mensch und Computer 2020 - Workshopband*, Bonn, 2020. doi: 10.18420/muc2020-ws105-375.
- [42] A. Hinderks, M. Schrepp, J. Thomaschewski, "UEQ-S Data Analysis Tool." https://www.ueq-online.org/Material/Short_UEQ_Data_Analysis_Tool.xlsx, 2024. [Online; accessed 24-April-2024].
- [43] A. Hinderks, M. Schrepp, J. Thomaschewski, "A benchmark for the short version of the user experience questionnaire," in *Proceedings of the 14th International Conference on Web Information Systems and Technologies (WEBIST 2018)*, 2018, pp. 373–377, SCITEPRESS - Science and Technology Publications. doi: 10.5220/0007188303730377.
- [44] A.-L. Meiners, M. Schrepp, A. Hinderks, J. Thomaschewski, "A benchmark for the UEQ+ framework: Construction of a simple tool to quickly interpret UEQ+ KPIs," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 9, no. 1, pp. 104–111, 2023, doi: 10.9781/ijimai.2023.05.003.
- [45] A. Hinderks, M. Schrepp, F. J. Domínguez Mayo, M. J. Escalona, J. Thomaschewski, "UEQ KPI value range based on the UEQ benchmark." doi: 10.13140/RG.2.2.34239.76967.
- [46] A. Hinderks, A.-L. Meiners, F. Mayo, J. Thomaschewski, "Interpreting the results from the user experience questionnaire (UEQ) using importance-performance analysis (IPA)," in *Proceedings of the 15th International Conference on Web Information Systems and Technologies, WEBIST 2019, Setubal, PRT*, 2019, p. 388–395, SCITEPRESS - Science and Technology Publications, Lda. doi: 10.5220/0008366503880395.
- [47] A. Hinderks, F. J. Domínguez-Mayo, A.-L. Meiners, J. Thomaschewski, "Applying importance-performance analysis (ipa) to interpret the results of the user experience questionnaire (UEQ)," *Journal of Web Engineering*, 2020, doi: 10.13052/jwe1540-9589.1926.
- [48] A. Hinderks, "A lifecycle for user experience management in agile development." PhD Thesis, Universidad de Sevilla, Sevilla, 2021.
- [49] M. Schrepp, J. Thomaschewski, K. Aufderhaar, "Do women and men perceive user experience differently?," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 6, pp. 63–67, 2019, doi: 10.9781/ijimai.2019.03.005.
- [50] J. Kollmorgen, M. Schrepp, J. Thomaschewski, "Influence of demographic variables and usage behaviour on the perceived user experience," in *Web Information Systems and Technologies*, Cham, 2023, pp. 186–208, Springer Nature Switzerland. isbn: 978-3-031-43088-6.
- [51] J. Kollmorgen, M. Schrepp, J. Thomaschewski, "Impact of usage behaviour on the user experience of netflix, microsoft powerpoint, bigbluebutton and zoom," in *International Conference on Web Information Systems and Technologies*, 2022, pp. 297–406, doi: 10.5220/0011380100003318.
- [52] H. Santoso, M. Schrepp, A. Hinderks, J. Thomaschewski, "Cultural differences in the perception of user experience." *Mensch und Computer 2017 - Tagungsband*, Regensburg, 2017. doi: 10.18420/muc2017-mci-0272.
- [53] M. Schrepp, H. Santoso, "Has culture an impact on the importance of ux aspects?," in *Mensch und Computer 2018 - Workshopband*, Bonn: Gesellschaft für Informatik e.V., 2018, doi: 10.18420/muc2018-ws04-0473.
- [54] H. B. Santoso, M. Schrepp, "The impact of culture and product on the subjective importance of user experience aspects," *Heliyon*, vol. 5, no. 9, p. e02434, 2019, doi: <https://doi.org/10.1016/j.heliyon.2019.e02434>.
- [55] A.-L. Meiners, J. Kollmorgen, M. Schrepp, J. Thomaschewski, "Which ux aspects are important for a software product? importance ratings of ux aspects for software products for measurement with the UEQ+," in *Proceedings of Mensch Und Computer 2021, Muc '21*, New York, NY, USA, 2021, p. 136–139, Association for Computing Machinery. doi: 10.1145/3473856.3473997.
- [56] A.-L. Meiners, A. Hinderks, J. Thomaschewski, "Trust, perspicuity, efficiency: Important ux aspects to consider for the successful adoption of collaboration tools in organisations," in *Computer-Human Interaction Research and Applications*, Cham, 2023, pp. 143–162, Springer Nature Switzerland. isbn: 978-3-031-49425-3.
- [57] M. Schrepp, "Measuring user experience with modular questionnaires," in *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2021, pp. 1–6. doi: 10.1109/ICACSIS53237.2021.9631321.
- [58] D. Winter, C. Hausmann, A. Hinderks, J. Thomaschewski, "Development of a shared ux vision based on ux factors ascertained through attribution," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 2, pp. 247–254, 2023, doi: 10.9781/ijimai.2023.04.001.
- [59] A. Hinderks, D. Winter, F. J. D. Mayo, M. J. Escalona, J. Thomaschewski, "Ux poker: Estimating the influence of user stories on user experience in early stage of agile development," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 7, pp. 97–104, 2022, doi: 10.9781/ijimai.2022.11.007.
- [60] A. Hinderks, F. J. D. Mayo, M. J. Escalona, J. Thomaschewski, "Requirements for user experience management - a tertiary study," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 6, pp. 160–167, 2024, doi: 10.9781/ijimai.2024.01.004.



Jessica Kollmorgen

Jessica Kollmorgen received a B.Sc. degree in Business Informatics, with parallel employment in backend software development, and a Master's degree in Media Informatics with a specialization in Mobile Computing and Security at the University of Applied Sciences Emden/Leer, Germany. Currently, she is a part-time research associate in the area of Agile User Experience. She has been an active member of the research group 'Agile Software Development and User Experience' since 2021. Her research interests lay on user experience factors, UX measurement and statistics as well as Agile UX.



Andreas Hinderks

Andreas Hinderks holds a Ph.D. in Computer Science by University of Seville (Spain). He became a Full Professor at the University of Applied Sciences and Arts Hannover (Germany) in March 2024. He has worked in various management roles as a Business Analyst and a programmer from 2001 to 2016. His focus lays on developing user-friendly business software. Currently, he is a freelancing Product Owner, Business Analyst and Senior UX Architect. He is involved in research activities dealing with UX questionnaires, measuring user experience and user experience management since 2011.



Jörg Thomaschewski

Jörg Thomaschewski received a Ph.D. in physics from the University of Bremen (Germany) in 1996. He became a Full Professor at the University of Applied Sciences Emden/Leer (Germany) in September 2000. His teaching and research focus is on Human-Computer Interaction, UX Management, Agile Software Development, and Requirements Engineering. In 2012 he founded the research group 'Agile Software Development and User Experience'. Dr. Thomaschewski has extensive experience in user experience training, UX questionnaires, agile methods, IT analysis, and consulting.

Platform for Improving the User Experience in the Creation of Educational Multiplayer Video Games

Fernando Sánchez Canella, Jordán Pascual Espada*, Irene Cid Rico

Department of Computer Science University of Oviedo (Spain)

* Corresponding author: jordansoy@gmail.com

Received 24 April 2023 | Accepted 30 November 2023 | Early Access 7 March 2024



ABSTRACT

Students' motivation is one of the factors that directly affect academic performance. In recent years, teachers are looking for ways to motivate students during their training period. For example, making use of slides, videos, films, comics or games to increase students' motivation to improve their learning experience. Some research works have revealed that multiplayer games which include cooperation and competition, among other factors, are an extraordinary tool for enhancing students' motivation. Current alternatives make it very complex for teachers to create multiplayer games for their students. The definition of the game requires many configurations and even technical knowledge. This research proposes a new platform that allows teachers to create multiplayer video games in a simple and fast way, improving the game creation process over current alternatives. The resulting games are also designed for to improve the student experience, and make it fun. These games do not only include trivia questions, but also use functional mechanisms from video games. The design of the generated games allows students to master the games in a short period of time during their classes.

KEYWORDS

E-learning, Gamification, Multimedia, Usability, Videogames.

DOI: 10.9781/ijimai.2024.03.002

I. INTRODUCTION

FROM an early age, we use games as a tool to teach and train students in different areas. Traditionally, as children grow up and were trained in more complex areas, these types of games decreased and were replaced by other teaching methods. In recent years the advancement of these technologies has boosted the increased use of games applied to training in different forms such as simulators, video games or other types of training environments [1]. In the educational field, games can be used to provide greater immersion in the subject matter being learned to [2], increase motivation [3], [1] improve satisfaction [4], increase entertainment, creativity or autonomy [5]. In some cases, even commercial videogames have been used in teaching [6]. We can establish different classifications of the games applied in education. On the one hand, there are video games created to deal with specific content. On the other hand, quiz-type video games provide teachers with a greater possibility to customize or configure the content.

There are other types of video games that are purely educational, since they have been created with the purpose of making players learn. We can consider these types of applications to be video games if they include the characteristics of video games. Depending on the researcher, there may be a fine difference between a video game and an application that applies gamification. The difference between

an application and a videogame is that the video game places the player in a virtual environment using 2D or 3D graphic resources. By gamification, we mean the inclusion of typical elements of a game to something that is not a videogame to motivate the people involved in the activity [7]. In practice, there can be a big difference between using gamification in the classroom and using a videogame as part of the educational process.

The effect of the increase in student motivation derived from the use of quiz game creation platforms is proven in the studies analyzed in the related work. The quiz platforms are increasingly used by teachers, due to several factors:

- They can be applied to almost any subject or content.
- The preparation time for the content is reasonable.
- The level of knowledge required by the teacher to configure the games is low.

In contrast, we believe that the level of motivation enhancement in a pure quiz game will not always be as high as in a more "traditional" video game, partly because quiz games are so simple that they do not include many of the features that positively impact students' motivation [8].

Video games used in education can include several factors that positively impact motivation [8]. Some of these features have been pointed out in several research works.

Please cite this article in press as:

F. Sánchez Canella, J. Pascual Espada, I. Cid Rico. Platform for Improving the User Experience in the Creation of Educational Multiplayer Video Games, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 4, pp. 140-147, 2025, <http://dx.doi.org/10.9781/ijimai.2024.03.002>

- **Internal interaction** among players, so that teams can be established or they can encourage cooperation [9], [10], [11].
- **Synchronization** between players so that they perform synchronous or asynchronous actions in the same scenario.
- **Roles** created to facilitate iteration and dependencies between players, e.g. a doctor, a builder, etc. [12]
- **Resources** (collectible objects in the game) can be finite or non-finite, consumable or non-consumable. These objects should be related to the educational context of the game [13].
- **Scores**, including a quantitative scoring system that motivate and stimulate competition [14].
- **Challenges**, there must be clear objectives that players have to meet and challenges that get progressively difficult or little repetitive, so that the results cannot be predicted [15], [16].
- **Rewards** (a way to stimulate players). Rewards normally make you earn points in a ranking or receive some distinction, objects or badges that provide satisfaction to students.
- **Artificial intelligence**, contributes positively to the immersion in the game.
- **Interoperability** (consideration of hardware requirements), so that the game can be used in an agile way in an educational environment.

Some researchers highlight several important factors to increase students' motivation. One of the factors is the possibility of establishing collaboration and competition among students [8], [11].

One of the challenges in using real video games in education is the complexity involved in defining their functionality, such as internal interaction, roles, scores, challenges, and rewards. This can make it difficult for teachers to configure video games for classroom use.

This research aims to address this challenge by creating a solution that allows teachers to quickly and easily define a video game with the key elements that have a positive impact on student motivation, as identified in previous research works. The focus is on improving the user experience for teachers in the steps of video game creation process. Additionally, the research seeks to enhance the experience of students when playing video games, ensuring that they can quickly master it within the limited time available in the classroom. After all, a game that is too complex for students to handle is of no use in an educational setting.

II. RELATED WORK

Since technologies are more present and more accessible among the population, the use of these in different fields has been investigated. One of the benefits of this technological development has been in the field of education, which has been used in combination with other educational resources.

Studies have been conducted using video games, analyzing how they can affect the learning process and student motivation. Some of these research works analyze how users are affected by the way objectives are set in the game. Some objectives are set to focus on learning, whereas others focus more on the completion of a specific task [17].

There are research works which analyze how group sizes affect students when they are using competitive multiplayer video games as a means of learning. It also analyzes how learners' cognitive load may increase as the group size increases [10].

Both studies mentioned above have used Minecraft's video game to conduct their analyses. The students' objective was to learn the basics of logic and programming. Evaluations were aimed at acquiring this type of knowledge and focused mainly on studying specific aspects

of learning. This game has a version called "Minecraft Edu" aimed at educational organizations so that teachers can develop their own work environments where they can teach subjects such as computer science, project management and environmental sustainability, among others.

Some studies show how gamification strategies in games affect the teaching of specific concepts [18]. This research exposes how certain elements within the games can positively affect the motivation and the perception of the activity by the students. They also expose the difficulty or lack of frameworks which allow teachers to design activities contained in games.

Many video games have been designed to cover very specific knowledge, competencies or objectives. For example, we can find video games for muscle rehabilitation [19] to simulate scientific applications [2] or to learn specific skills such as how to fly airplanes through simulation video games [20]. Thanks to features such as 3D graphics, music, animations and iteration capacity, video games can achieve a higher immersion than other more classical methods.

There are some educational video games that have editors and tools for teachers to configure and edit the content. One of the most famous video games in this category is Minecraft Education [21], where teachers can create maps, mechanics, objects, puzzles and questions, among other things. This version of Minecraft has been used successfully in numerous cases. For example, this was used by more than 100 students in a university environment, to reinforce programming and logic knowledge. A positive aspect of Minecraft Education is that it can be multiplayer and allow competition among students [10].

There are platforms such as Scratch and Roblox, which can be used for developing video games. These do not require in-depth knowledge of programming, 3D modelling or videogame design. An example of Scratch used in the educational field is shown in a study on the subject of English with students with Attention Deficit Hyperactivity Disorder (ADHD) [22]. In this case, Scratch was used to create several mini games which introduced concepts of different topics applied to learning English [23].

Multiple platforms allow teachers to create quiz games. These games include various gamification elements, but they may not be considered real video games and the level of motivation enhancement may be lower compared to other video games. Kahoot is one of the most popular quiz game creation platforms [24]. This tool consists of exposing questions previously configured by the teacher to the students. In each question, several possible answers are exposed, and students individually must select the answer they consider correct. In the case of selecting the correct answer, the game assigns you a series of points depending on the speed with which you have answered the question. Between each question, a ranking is shown to the students which generates a multiplayer and competitive environment. There are many other relatively similar platforms such as Quizizz [25]. Several pieces of research prove that the use of this type of game has had very positive effects on the motivation of students. 90% of students who use it occasionally or frequently reported that they had fun using the system and more than 80% would like to use it in other subjects [26]. Other studies indicate that the use of these tools increased the interest of students by more than 60% [27].

Only a few tools or proposals that enable the creation of video games or educational video games include features that can enhance student motivation. Many so-called educational games are simply quiz applications with some gamification aspects, lacking actual game features or the elements that have been identified in numerous research studies as being able to enhance student motivation. While it is relatively easy for teachers to create questions using such platforms, the resulting products are not genuine video games.

The limited alternatives that allow for the creation of actual video games, such as Minecraft EDU, can be challenging for teachers as they require technical expertise and can make the process of creating a video game complex and time-consuming. Therefore, there is a significant need to create new solutions that can enable teachers to create engaging video games that incorporate the key elements of motivation identified in research studies, but without the technical challenges associated with existing tools. This would enable teachers to create high-quality educational video games that promote student engagement and learning in a fun and interactive way, without requiring significant technical knowledge. No platform or proposal has been found that focuses on improving the experience of creating multiplayer educational games.

III. PROPOSED PLATFORM

We propose a platform focused on improving the user experience of teachers in creating competitive multiplayer videogames. With this platform, teachers can configure the game and integrate the questions related to the current session in an agile way. Each game session is configured with a group of questions about a subject. These questions have a similar structure to the questions which are used in applications such as Kahoot, but in this case, they will be integrated within a real video game with game mechanics and rules (Fig. 1).



Fig. 1. Proposal conceptual scheme.

Another point of interest is the output game created by teachers. The research sets several points to consider in output game design, thinking about increasing the motivation of the students. It has been decided to base the initial idea on the popular game BedWars, a conventional Minecraft mini game in which players must destroy their rivals' beds. The original concept and rules of the game have been combined and adapted to fill the research goals.

It is primarily a real-time multiplayer game. At the beginning of the game, each student starts on his own island. On each island, there is a "block generator" and a "flag." Initially, the islands are separated by a sea. The goal of each student is to move to other islands to capture their rivals' flags and then take the flags to their own islands. The player who captures as many flags as he has set on the game configuration will win the game. To move between the different islands, players must build bridges using blocks. The blocks are collected in the block generators.

The output game has been designed to prioritize the students' user experience. In many instances, students may only have a limited amount of time to play, such as a single class period. Consequently, the game's controls and mechanics must be intuitive, enabling students to quickly become proficient without spending excessive time familiarizing themselves with the gameplay. The design of the gaming experience aims to facilitate the mastery of the videogame, even for those without prior experience.

In addition to the students' user experience, the game must continue to maintain features identified as beneficial for improving user motivation.

Internal interaction among players. The game is designed to keep all players visible on the same map. The game mechanics will also offer the chance to collaborate and compete with other players. The players may also use elements created by other players (bridges).

Roles. The proposed design does not fully include this feature, i.e. there are no players with their own skills which cannot be obtained by another player. The use of roles has as its main objective that students may need the skills of others. In this game, all players have the same skills, and can build bridges if they have enough blocks. In some instances, there could be some players who cannot build bridges, which will help encourage collaboration.

Resources This guideline recommends the use of finite or non-finite resources, consumable or non-consumable and related to the educational context of the game. In this case, the main resource are the building blocks. Blocks are a finite resource that is created by the generators progressively with several new blocks created with each turn. The blocks are a resource, as they can be used to create bridges, which is an essential element of the game. Without bridges, it is not possible to move to other islands. The resources are directly related to the educational context since a question is sent to the student when they request the blocks of a generator. If answered correctly, the student will get some extra blocks.

The design includes a quantitative scoring system, with the objective to stimulate competition among the students. At any time during the game, it is possible to see the number of points that each participant has. In this case, the points are the number of flags that have been captured on other islands and then brought back to the player's island.

The proposed challenges are clear objectives, which the players must fulfill at each moment. In an optimal scenario, these challenges should be non-repetitive and increase their level of complexity. The main challenge is clearly identified - to advance towards a rival flag in order to capture it. This challenge involves carrying out a series of tasks, such as obtaining the building blocks and creating the bridge to reach the other player's island. The fewer flags that are available, the more complex it will be to obtain them, since there are a greater number of players who will go for them. In general, multiplayer games prevent challenges from being presented in a very repetitive way as there are several people involved. Another secondary challenge that the player could face in the game is to "defend" his own flag. Through A game feature that has been included that allows the player to challenge other players to a question, which can be used to prevent them from attacking the player's flag.

The rewards are encouragements that allow players to gain ranking points or recognition. The proposal includes two types of rewards (1) the reward for saving a flag that consists in getting a point for the ranking (2) the reward for answering questions that results in getting extra building blocks.

The current design does not include **Artificial Intelligence**, but research shows that its use can contribute positively to the immersion in the game. In this case, the proposed design is imminently multiplayer, so it is not so critical that non-controllable elements (NPCs) managed by artificial intelligences appear.

Interoperability, the platform has been designed on web technologies, so the hardware requirements are extremely light. The game only requires a web browser for its operation avoiding having to install a program and its subsequent updates and even opening the possibility for students to play on their own devices such as cell phones or tablets.

The output game is played in turns of limited time. In these turns, the player has the possibility to move a maximum distance and perform an action. There are different actions, and it is in these actions

that the questions are introduced (Fig. 2). The more questions that are answered correctly, the more competitive advantage the student will have during the game. The main actions are:

- **Move:** Each player can move 8 cells each turn. The floor of the map is divided into cells, and they can only move through cells containing land or bridges.
- **Collect blocks:** If the player moves next to a “block generator” he will be asked a question. He can get more blocks if he answers correctly.
- **Building bridges using blocks:** This feature is necessary to move from one island to another. The blocks placed disappear after a number of turns, with the aim of encouraging the obtaining of blocks and therefore answering questions on the subject.
- **Capture and deposit the flags:** By moving the player to the same square as the rival flag, a player is able to capture it. From that moment, the player owns the flag and must deposit it on his own island. Once it is deposited, they will get a point.
- **Challenge another player:** When two players are close to each other, either of them can initiate a challenge. Challenges send the same question to both players. In the event that either player does not answer correctly, they will be penalized by losing half of their blocks and being transported to their island.
- **Collaborate with another player:** When two students are close to each other, either of them can initiate a collaboration. In this case, the two players team up to answer a question together and cooperate. If one of the players answers the question correctly, both players will receive an extra number of blocks. In case they answer incorrectly, both would be penalized.

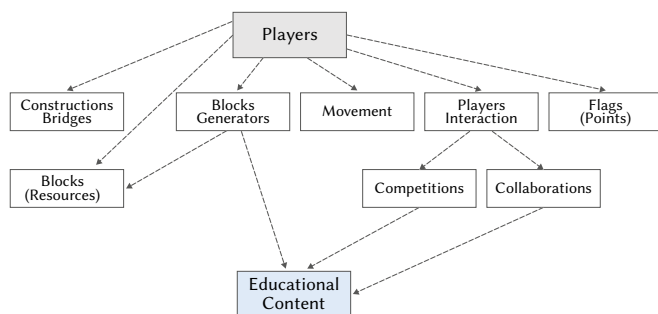


Fig. 2. Gameplay design of the game and tree of actions.

The responsibility for configuring and managing the game rests with the teacher, who must customize various aspects of the game to align with their teaching goals. To simplify the configuration process, our solution includes default parameters, which can be easily adjusted to meet the specific needs of the classroom. Additionally, the teacher can integrate their own questions into the game by specifying the statement and answer choices (true or false). These customizable features enable the teacher to tailor the game to their subject matter and classroom objective. It is possible to set a statement, a question and one or more correct answers. The system can be configured to force the student to write a text answer in a free field or to choose between different predefined options where only one is correct. In addition, the teacher must also enter the incorrect alternatives. The teacher has the possibility to set up several sessions with different questions and invite different students to each of them.

The platform consists of 3 subsystems.

- **Administration subsystem.** Where the teachers can create their games, modify the parameters of the default configuration, and add questions to the content. It is also the system in charge of organizing the students in different game sessions. The interfaces and parameters of this module have been studied in detail so that

teachers have the best possible user experience and are able to generate a game quickly and without errors during the process. The subsystem has been developed using the VueJs Framework and Firebase tools.

- **Game Client subsystem.** This consists of the video game prototype that has been designed. With a very simple graphic design of 2.5 dimensions, it is very light and interoperable in order to be used on low-powered computers and thus be available to the maximum number of students. This has been developed mainly in JavaScript to be used in a browser.
- **Server client subsystem.** This is in charge of synchronizing the video game between all the players and the platform data. In this way, the questions that have been configured during the game are included in the game. For this module, we have used technologies such as NodeJs on the server side, and WebSockets as a means to establish communication with the clients of the videogame.

IV. USE CASE

This use case provides a comprehensive example of how the platform can be used. To begin, the teacher accesses the platform via the web and defines the parameters of the game they wish to create. They can then enter the players and modify default game settings as desired. Once this has been completed, the teacher includes questions (Fig. 3) questions can be included in the editor or imported into different formats. The teacher invites players to join the game, and configures the minimum number of students required for the game to start. By following these steps, the teacher can seamlessly and efficiently create an engaging and interactive learning experience for their students.



Fig. 3. Configuration view.

Each student will join the platform. They can select a session on the main page. At the start of the game, each player is transported to his own island and must wait his turn. Each island has a flag in the middle and a block generator. The turn information can be found at the top left of the screen, and in it shows the number of flags and the building blocks available for the player. In the player's turn he can move using the mouse, with the movement being limited to a maximum number of cells.

When the player moves to a cell adjacent to a block generator, he can retrieve the blocks stored in the generator. The generators create blocks at every turn. The collection of blocks is one of the actions that has associated questions. When collecting the blocks a dialogue box will appear for the player to answer the question (Fig. 4). If the answer is correct, the player will receive twice as many blocks.

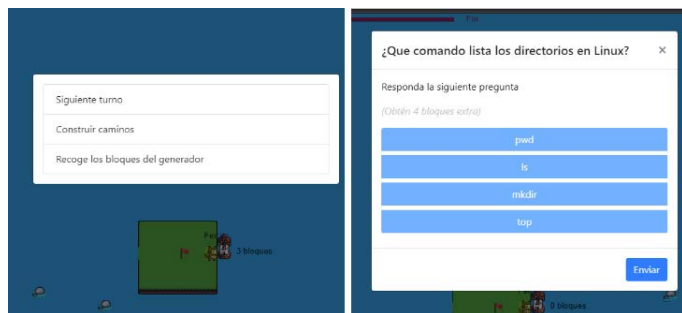


Fig. 4. Player answering educational question.

Players have to move to other islands, so initially they must collect blocks to build bridges to the other islands. The objective of moving to another island is to capture the rival flag. The construction of bridges is done in the same way as the movements, but instead of by land it is done by sea. When moving the mouse, the blocks will be discounted, and when clicking the mouse, the bridge will be built (Fig. 5). The bridges can be used by all players, not only by the one who built them. These bridges have a finite duration and begin a process of destruction once built as they disappear after 5 turns.



Fig. 5. Player building a bridge.

Once a player gets an opponent's flag, he will carry it over (Fig. 6). At that point, he must carry the flag to his own island to get a point. Once the player delivers the rival flag to his island, the flag reappears in its original place, opening the possibility for it to be captured again by another player. The ranking indicates the number of flags retrieved by each player. The teacher can set a number of flags to obtain or simply let the students play for a period of time and see how many they get.

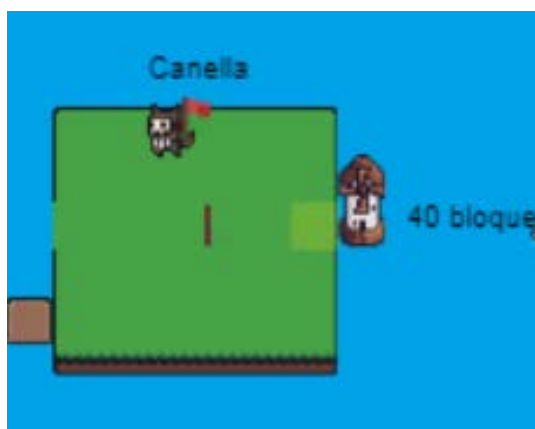


Fig. 6. Player getting an enemy flag.

V. EVALUATION

In order to validate the research goals, a three-phase evaluation process was carried out:

1. Evaluate the game creation process. Analyzing how long it takes to create a collaborative multiplayer game with the proposal and comparing it with current alternatives.
2. Evaluate the students' gaming interaction, focusing on new students who have never seen the game before to learn the mechanics quickly. On the test of the proposal carried out in the classroom, game logs have been obtained and the evolution in the number of actions performed by the students has been compared.
3. Impact on motivation. To evaluate the use of the output game and its impact on **motivation**, first, a comparison was made of the characteristics that potentially allow an increase in motivation, and the proposal was compared with the most popular alternative currently available. Secondly, the game was used in a real classroom and a survey was conducted asking students.

A. Game Creation Process

The incorporation of this type of video games in the classroom requires prior knowledge and effort on the part of the teacher. Therefore, in this evaluation we will try to estimate the amount of effort required by the teacher to prepare an activity with each of the alternatives. We set the objective of creating and sharing a simple multiplayer game which include a question and four options, which offers feedback to the student who answers the question.

In order to measure the complexity of creating this game, two evaluations were carried out - a first evaluation to measure the user's keyboard and mouse activity with the mousetron tool, and a second measurement based on the KLM-GOMS technique [28] to estimate the amount of time it takes an average user to interact to complete a specific task. All games have been created by an expert user who knew how to operate the tools. As no confusion or learning is being taken into account, this would be a near optimal iteration. This type of evaluation abstracts the human factor since the teacher's background can be very different.

The results show (Table I) how our proposal has a similar complexity to other alternatives just based in questions, and which are not real games. It is much simpler with respect to more complex platforms such as Minecraft or Roblox, which require 7.29 and 11.31 times the time needed in our prototype.

The evaluation based on the KLM-goms technique monitored five variables. The number of times a keyboard key was pressed (K). The number of times it was necessary to point to an object (P). The number of mouse clicks (B). The number of transitions made between the mouse and the keyboard (H). The number of times the user interface displayed an element on the screen that required extra mental preparation (M). The in-game action needed (G). Assigning each action a time in seconds of 0.20 (K), 1.10 (P), 0.10 (B), 0.40 (H), 1.20 (M) and 8.5 (G) respectively [28].

The times obtained with KML-Goms (Table II) show that Kahoot, Quizlet, Arcade Game Generator and our proposal are the most agile platforms, with times of around 25 seconds, Kahoot being the fastest with 20,8. Minecraft Edu obtained a time of 578 seconds, 154,2 seconds for Scratch and 805,8 seconds for Roblox. This points how the most flexible platforms when it comes to creating customized experiences also require a greater amount of work on the part of the user. The times measured with the KLM-goms procedure, are the times that the user interacts only with the user interface, while the Mousetron measures the total time of the task. That means that KLM-goms skip the loading times of each platform.

B. Students' Gaming Interaction

In order to determine the degree to which the game is easy to learn, the actions performed by the students in the test scenario have been recorded. The following graph (Fig. 7) shows the evolution of a

TABLE I. MOUSETRON METRICS

	Time (s)	Mouse movement	Key strokes	Mouse left button	Mouse right button	Double click	Wheel
Proposal Game	86	190.5	118	21	0	0	0
Kahoot	89	358.14	63	19	0	1	10
Quizlet	76	388.62	75	15	1	0	33
Arcade Game Generator	72	198.25	79	13	0	0	0
Minecraft EDU	627	1780.54	384	130	53	11	128
Scratch	235	2034.54	105	116	3	22	182
Roblox	973	16312.51	642	498	41	16	441

TABLE II. KLM-GOMS METRICS

	k	p	b	h	m	g	Total (s)
Proposal	27	7	5	3	9	0	25,6
Kahoot	26	4	4	3	8	0	20,8
Quizlet	27	6	4	3	8	0	23,2
Arcade Game Generator	32	6	5	4	9	0	25,9
Minecraft EDU	29	11	14	7	17	63	578
Scratch	52	17	20	11	21	11	154,2
Roblox	82	43	32	18	36	81	805,8

13-players game and the number of actions the students performed in each minute. The starting point was a game designed with questions about operating systems. The students played the game. As the game has a maximum game size of 12 players, they were randomly divided into 2, playing exactly the same game. The actions recorded do not include the movements.

The test performed had turns of 20 seconds, this time being extended when a student receives a question. In each turn, the students perform a movement and an action. When they finish the action, the turn ends. If no action is performed, the turn ends automatically after 20 seconds. The graph shows how, as the session progresses, the number of actions per second increases. It reaches up to 19 actions per minute. In the 22-minute period that the game lasted, an evolution in the speed of the game can be observed.

At the beginning of the test, values are below 5 actions. Turns were running out for some players. This is due to the fact that it was the first time that the students used the game, and they did not know the controls and mechanics of the game well, despite having been briefly explained at the beginning of the test. When the number of actions per minute exceeds 6, it means that all 6 players have been able to perform their action in one minute, which means that each player takes about 10 seconds, half the preset time in one turn. When the number of actions exceeded 12 the number of actions was quite fast indicating that they already had a very high level of understanding of the game.

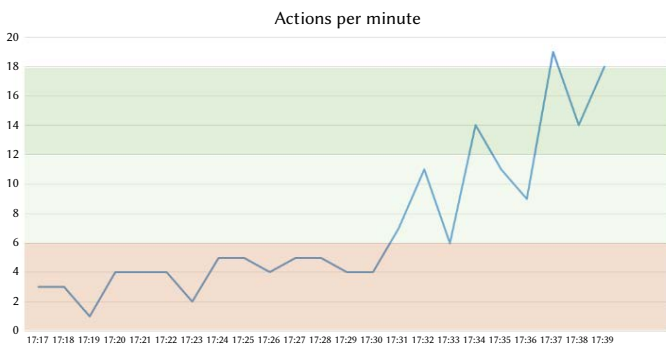


Fig. 7. Actions from students per minute.

C. Impact on Motivation

To evaluate the use of the output game and its impact on motivation, first, a comparison was made of the characteristics

that potentially allow an increase in motivation, and the proposal was compared with the most popular alternative currently available. At the designing of the game a series of characteristics were identified that were estimated to positively influence game motivation. We point out that the mere fact of including these features does not mean that the level of motivation is necessarily higher. We analyzed which of these features could be fulfilled by the proposal and other alternatives. Fig. 8 shows the results.

Generated game complies with 6 of the 8 characteristics evaluated. It needs an internet connection and a web browser. One of the most popular solutions, Kahoot, complies with only 3 features. Minecraft EDU and Roblox, could meet all the features depending on the implementation of the game by the teacher. It should be noted that both options require the installation of a client application. Popular platforms such as Scratch can be used to create video games in educational environments, but these video games cannot have synchronization between players as they are mainly designed for a single player. On the other hand, we find platforms such as Arcade Game that could only meet two of the evaluated characteristics.

The seven tools analyzed could be used to create game dynamics with subject content. The degree of customization that can be achieved with each tool varies, although greater customization requires a deeper knowledge of the platform or knowledge of programming concepts. Our proposal and tools such as Kahoot, Quizlet or Arcade Game generator offer very limited actions. Teachers must configure some values and introduce the subject information. On the other hand, Minecraft, Roblox and Scratch are game creation platforms that have different modalities and even make use of development tools, which require a higher level of knowledge.

Secondly, the game was used in a real classroom and a survey was conducted asking students. The solution was evaluated in a class with 13 students from a vocational training center, taking the subject of operating systems. Of which 77% were men, and 23% were women. 85% percent of the students acknowledged playing video games on a regular basis.

Eighty-six questions were prepared with four possible answers. Each question had only one correct option. These questions have two types of design, a first design where a statement is exposed, in which a word is missing, and students must complete the sentence, and another type where a question is directly posed and the correct answer must be chosen.

They are used to using Kahoot in their classes and they know how it works. They have also recently used Kahoot in a class. This is relevant because we are going to ask them in the survey if they would like to use the proposal or Kahoot more. Kahoot works like a trivia game where a question is presented to all students at the same time and they have a time limit to answer. If students answer the question correctly, they will get points. The faster they answer the question correctly, the more points the students get. After each question, they are shown a ranking of the students' scores.

The proposed solution was used during a class session. Initially, the basic functioning of the game was explained to the students for 10 minutes. The students were divided into two groups of 7 and 6 students. Each of the groups played in one server. We use this because there is a limit of 12 players per map, and both groups played exactly the same game. They played a total of two games for a total of 25 minutes.

Once the game session was over, a survey was carried out with a series of questions to analyze the impression they had had during the session.

- **Q1** *Do you think that thanks to the game, your level of motivation has been higher than other classes?*

Students could choose between the options "much lower, lower, higher, much higher". In this case, 46.2% of the students answered that the experience with respect to other classes had been "Much higher," 46.2% answered "Higher" and only 7.7% "much lower".

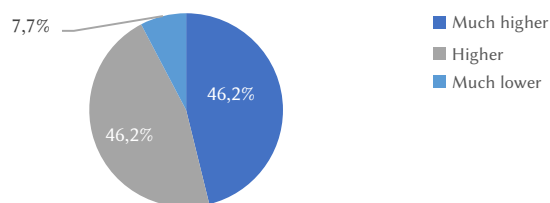


Fig. 8. Responses percentage.

On the other hand, we compared the experience using Kahoot, with our proposal that integrates questions within a competitive video game. For this purpose, the following question was posed:

- **Q2** *What type of game do you prefer?*
 - A. "A game of only questions based on classroom content (Kahoot type)."
 - B. A multiplayer video game in which answering questions can provide some competitive advantage in the game (video game used in the session)

On this question, 92.3% of the students chose the second option. This showed that their experience had been more positive.

We were also interested to know if the students preferred to do these types of review activities together with their classmates creating a dynamic of competition, or if they preferred to do a similar activity individually. In order to know their opinion, we asked them the following question.

- **Q3** *To what degree do you consider that doing these types of review exercises together with your peers is more satisfying than doing them individually?*

They could answer by choosing from a range of 1 to 5, with 1 being "Less satisfactory" and 5 being "More satisfactory". The average obtained in the students' answers was 4.46 with a deviation of 0.78.

In order to find out to what degree they had found the experience positive and believe that it would be useful to transfer these types of activities to other subjects, they were asked:

- **Q4** *Would you recommend using this game to other teachers?*

They were able to choose on a scale of 1 to 5 both included. The lowest score means that they would not recommend it and a 5 means that they would. The responses obtained an average of 4.46 and a standard deviation of 0.88.

Finally, they were also given the opportunity to make some comments about their experience with the game. And how they think it affected their motivation or how useful they think it can be in their learning process. Some comments highlight the usefulness of the application to set concepts or memorize concepts that they needed for the subject. They also pointed out how positive it had been for them to carry out this type of activity after several hours of class.

VI. CONCLUSIONS AND FUTURE WORK

In this research work, we focused on providing an innovative solution that enhances the user experience of teachers when creating multiplayer educational games. We observed that most common educational games, like Kahoot, lack real game mechanics and mainly consist of questions with some gamification features. On the other hand, real multiplayer games, such as Minecraft EDU, require a significant amount of work from teachers to define and configure the game's parameters.

Our proposal allows teachers to create multiplayer cooperative video games with real video game features without the need for extensive configuration. Based on our evaluation, teachers can easily configure their game and teaching content. In comparison to other tools such as Kahoot or Quizlet, our proposal's complexity is similar, with a minimal difference of only 3 seconds in time required for usage. Furthermore, our proposal significantly outperforms other alternatives such as Minecraft EDU or Roblox, which require up to 700% more time for configuration.

The games generated through our proposals feature user-friendly video game mechanics that are easy for students to understand and control. To evaluate this objective, we measured the number of actions performed by students who tested the game. Results showed that within approximately 14 minutes, students were able to play the video game smoothly at an expert player's level.

Additionally, we assessed student motivation compared to the reference Gamification tool Kahoot by distributing questionnaires to the students participating in the test. The response to the proposed solution was overwhelmingly positive, with 92.3% of students preferring it over tools based on questions, such as Kahoot. Furthermore, 46.2% of students reported in the survey that thanks to the game, the level of motivation was much higher than in other classes.

Students were highly satisfied with the opportunity to work collaboratively on questions (4.46/5) and most recommended the use of the proposal in other subjects (4.46/5). Overall, the proposal's impact on motivation and engagement in the classroom was remarkable, and it was well received by students as a valuable and exciting tool for learning.

Overall, our innovative solution offers a streamlined and efficient process for teachers to create multiplayer educational games with real game mechanics and engaging features, ultimately enhancing the learning experience for students.

As future work, we are considering the possibility of looking for systems that automate the generation of questions or content that can be implemented in this type of video games. For the preparation of the sessions, extra work has been required by the teachers to prepare the questions and configure the activity within the platform. Therefore, we are interested in evaluating whether the use of this type of tools achieves greater acceptance by the teachers, without losing quality of the learning content.

REFERENCES

- [1] C. González-González and F. Blanco-Izquierdo, "Designing social videogames for educational uses," *Computers and Education*, vol. 58, no. 1, pp. 250–262, Jan. 2012, doi: 10.1016/j.compedu.2011.08.014.
- [2] S. A. Barab, B. Scott, S. Siyahhan, R. Goldstone, A. Ingram-Goble, S. J. Zuiker, and S. Warren, "Transformational Play as a Curricular Scaffold: Using Videogames to Support Science Education," *Journal of Science Education and Technology*, vol. 18, no. 4, p. 305, 2009, doi: 10.1007/s10956-009-9171-5.
- [3] A. Tlili, M. Chang, J. Moon, Z. Liu, D. Burgos, N. Chen, and Kinshuk, "A Systematic Literature Review of Empirical Studies on Learning Analytics in Educational Games," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 2, pp. 250–261, Dec. 2021, doi: 10.9781/ijimai.2021.03.003.
- [4] M. C. Ramos-Vega, V. M. Palma-Morales, D. Pérez-Marín, and J. M. Moguerza, "Stimulating children's engagement with an educational serious videogame using Lean UX co-design," *Entertainment Computing*, vol. 38, p. 100405, May 2021, doi: 10.1016/j.entcom.2021.100405.
- [5] Z. Zainuddin, M. Shujahat, H. Haruna, and S. K. W. Chu, "The role of gamified e-quizzes on student learning and engagement: An interactive gamification solution for a formative assessment system," *Computers & Education*, vol. 145, p. 103729, Feb. 2020, doi: 10.1016/j.compedu.2019.103729.
- [6] N. P. Zea, *Metodología para el diseño de videojuegos educativos sobre una arquitectura para el análisis del aprendizaje colaborativo*. Universidad de Granada, 2011.
- [7] D. Dicheva, C. Dichev, G. Agre, and G. Angelova, "Gamification in education: A systematic mapping study," *Educational Technology & Society*, vol. 18, no. 3, pp. 75–88, 2015.
- [8] D. Buchinger and M. da Silva Hounsell, "Guidelines for designing and using collaborative-competitive serious games," *Computers & Education*, vol. 118, pp. 133–149, Mar. 2018, doi: 10.1016/j.compedu.2017.11.007.
- [9] A. Tlili, S. Hattab, F. Essalmi, N. Chen, R. Huang, Kinshuk, M. Chang, and D. Burgos, "A Smart Collaborative Educational Game with Learning Analytics to Support English Vocabulary Teaching," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 215–224, May 2021, doi: 10.9781/ijimai.2021.03.002.
- [10] S. Nebel, S. Schneider, and G. D. Rey, "From duels to classroom competition: Social competition and learning in educational videogames within different group sizes," *Computers in Human Behavior*, vol. 55, pp. 384–398, Feb. 2016, doi: 10.1016/j.chb.2015.09.035.
- [11] J. Sánchez-Martín, F. Cañada-Cañada, and M. A. Dávila-Acedo, "Just a game? Gamifying a general science class at university: Collaborative and competitive work implications," *Thinking Skills and Creativity*, vol. 26, pp. 51–59, Dec. 2017, doi: 10.1016/j.tsc.2017.05.003.
- [12] Z.-H. Chen, C.-Y. Chou, Y.-C. Deng, and T.-W. Chan, "Animal Companions as Motivators for Teammates Helping Each Other Learn," in *Proceedings of Th 2005 Conference on Computer Support for Collaborative Learning: Learning 2005: The next 10 Years!*, 2005, pp. 43–47.
- [13] M. E. Gorman, "Serious games, sustainable civilizations and trading zones," in *2009 IEEE International Symposium on Sustainable Systems and Technology*, 2009, pp. 1–3, doi: 10.1109/ISSST.2009.5156702.
- [14] R. Takaoka, M. Shimokawa, and T. Okamoto, "A Development of Game-Based Learning Environment to Activate Interaction among Learners," *IEICE Transactions on Information and Systems*, vol. E95.D, pp. 911–920, Apr. 2012, doi: 10.1587/transinf.E95.D.911.
- [15] J. Ribelles, A. López, and V. J. Traver, "Modulating the Gameplay Challenge Through Simple Visual Computing Elements: A Cube Puzzle Case Study," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 6, pp. 177–193, 2024, doi: 10.9781/ijimai.2022.05.001.
- [16] D. Liu, X. Li, and R. Santhanam, "Digital Games and Beyond: What Happens When Players Compete?," *MIS Quarterly*, vol. 37, no. 1, pp. 111–124, 2013.
- [17] S. Nebel, S. Schneider, J. Schledjewski, and G. D. Rey, "Goal-Setting in Educational Video Games: Comparing Goal-Setting Theory and the Goal-Free Effect," *Simulation & Gaming*, vol. 48, no. 1, pp. 98–130, Nov. 2016, doi: 10.1177/1046878116680869.
- [18] J. C. Díez Rioja, D. Bañeres Besora, and M. Serra Vizern, "Experiencia de gamificación en Secundaria en el Aprendizaje de Sistemas Digitales= Gamification Experience in Secondary Education on Learning of Digital Systems," *Experiencia de gamificación en Secundaria en el Aprendizaje de Sistemas Digitales= Gamification Experience in Secondary Education on Learning of Digital Systems*, pp. 85–105, 2017.
- [19] G. Saposnik, M. Mamdani, M. Bayley, K. E. Thorpe, J. Hall, L. G. Cohen, and R. Teasell, "Effectiveness of Virtual Reality Exercises in STroke Rehabilitation (EVREST): Rationale, Design, and Protocol of a Pilot Randomized Clinical Trial Assessing the Wii Gaming System," *International Journal of Stroke*, vol. 5, no. 1, pp. 47–51, Feb. 2010, doi: 10.1111/j.1747-4949.2009.00404.x.
- [20] Q. Kennedy, J. L. Taylor, G. Reade, and J. A. Yesavage, "Age and expertise effects in aviation decision making and flight control in a flight simulator," *Aviation Space and Environmental Medicine*, vol. 81, no. 5, pp. 489–497, 2010.
- [21] G. Ekaputra, C. Lim, and K. I. Eng, "Minecraft: A game as an education and scientific learning tool," *The Information Systems International Conference (ISICO) 2013*, vol. 2013, 2013.
- [22] D. L. Aldana-Avilés, "El lenguaje de programación Scratch como material didáctico motivador para la aplicación y evaluación de contenidos en el área de inglés para alumnos con diagnóstico de TDAH," 2015.
- [23] C. Meier, J. Saorín, A. B. de León, and A. G. Cobos, "Using the Roblox Video Game Engine for Creating Virtual tours and Learning about the Sculptural Heritage," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 15, no. 20, pp. 268–280, 2020.
- [24] M. A.-A. Ismail and J. A.-M. Mohammad, "Kahoot: A promising tool for formative assessment in medical education," *Education in Medicine Journal*, vol. 9, no. 2, 2017.
- [25] D. Orhan Göksün and G. Gürsoy, "Comparing success and engagement in gamified learning experiences via Kahoot and Quizizz," *Computers & Education*, vol. 135, pp. 15–29, Jul. 2019, doi: 10.1016/j.compedu.2019.02.015.
- [26] A. I. Wang, "The wear out effect of a game-based student response system," *Computers & Education*, vol. 82, pp. 217–227, Mar. 2015, doi: 10.1016/j.compedu.2014.11.004.
- [27] M. Martín-Sómer, J. Moreira, and C. Casado, "Use of Kahoot! to keep students' motivation during online classes in the lockdown period caused by Covid 19," *Education for Chemical Engineers*, vol. 36, pp. 154–159, Jul. 2021, doi: 10.1016/j.ece.2021.05.005.
- [28] D. Kieras, "Using the Keystroke-Level Model to Estimate Execution Times," Mar. 2003.



Fernando Sánchez Canella

Is an Engineering in Computer Systems and M.S. in Web Engineering from School of Computer Engineering of Oviedo (University of Oviedo, Spain). His research interests are in the field of the multimedia web applications, videogames, graphic applications and the use of computers for education.



Jordán Pascual Espada

Is an associate Professor at Computer Science Department of the University of Oviedo. Ph.D. from the University of Oviedo in Computer Engineering. His research interests include graphic applications, videogames the Internet of Things, exploration of new applications and associated human computer interaction issues in ubiquitous computing and emerging technologies, particularly mobile and Web.



Irene Rico Cid

Is an assistant Professor at Computer Science Engineering in Computer Systems and M.S. Computer Science from School of Computer Department of the University of Oviedo. Currently, she is a Ph.D candidate. Her research interests are in the field of the multimedia web applications, videogames, graphic applications and the use of computers for education.

