UNIR LA UNIVERSIDAD EN INTERNET

"We can only see a short distance ahead, but we can see plenty there that needs to be done."
Alan Turing

EDITORIAL TEAM

Editor's Note

THE International Journal of Interactive Multimedia and Artificial Intelligence (IJIMAI) is a diamond open access journal which provides an interdisciplinary forum in which scientists and professionals can share their research results and report new advances on artificial intelligence tools, theory, methodologies, systems, architectures integrating multiple technologies, problems including demonstrations of effectiveness, or tools that use AI with interactive multimedia techniques. The journal is supported by Universidad Internacional de La Rioja (UNIR) and by all those members of this multicultural community who, with a sense of commitment to the development of science, dedicate their knowledge and time to authoring, editing and reviewing tasks, and without whom this knowledge sharing project would not be possible.

This regular issue begins with a series of five articles covering key advancements in the area of computing vision. The first three propose solutions that are applied in the health and welfare fields. Specifically, the first one targets the diabetic foot ulcers (DFU), which are among the most serious diabetic complications. Kumar Das et al. propose a federated learning-based solution to automatically diagnose DFU from patient images, addressing key common challenges in healthcare applications as data privacy or diagnostic accuracy. With a decentralized learning proposal, machine learning models are trained directly on client devices, where the sensitive patient data is kept, avoiding the privacy concerns that arise in centralized systems. The solution uses a hybridized data augmentation technique together with a lightweight convolutional neural network (CNN) running on resource constrained devices, which shows good performance. Therefore, the proposal has the potential to provide accessible, affordable and privacy-preserving diagnostic support of DFU, even in developing regions.

The research by Irfan et al. focuses on Alzheimer's disease, a common type of dementia that is expected to affect more and more people in the coming years. This research aims to deploy deep learning methods to determine if they can extract helpful Alzheimer's disease biomarkers from magnetic resonance imaging (MRI) and classify brain images into Alzheimer's disease, mild cognitive impairment, and normal cognitive groups. Specifically, various CNN are trained to predict Alzheimer's disease using different views of MRI images, including sagittal, transverse, and coronal views. Authors propose an intelligent probabilistic approach to select slice numbers for the three MRI views to reduce computational cost. With hyperparameter tuning, batch normalization, intelligent slice selection and cropping, and combination of the views, an accuracy of 92.21% is achieved, showing better performance than related studies.

To improve the quality of life of the elderly, Liu et al. propose a fall detection scheme based on human skeleton nodes that could facilitate real-time fall detection, with the corresponding needed immediate help. In their research, a hybrid model based on spatial-temporal graph convolutional network (ST-GCN) and YOLO algorithm is proposed for multi-person fall detection. The first network is used to detect the fall action, while the second is used for accurate and fast recognition of multi-person targets. Optimization methods are also used to achieve real-time performance. Authors use both public single-person datasets and their own multi-person dataset in their experiments. They find their proposal has high detection accuracy under better environmental conditions, compared to state-of-the-art schemes, and outperforms other models in terms of inference speed.

Next article by Su et al. presents a solution to detect human-object interactions (HOI) in images. HOI detection goes a step further in object detection in computer vision, detecting the human and the object and extracting the semantic relationship between both. The authors propose a spatial-aware multilevel parsing network (SMPNet) that uses a multi-level information detection strategy, including instance-level visual features of detected human-object pair, part-level related features of the human body, and scene-level features extracted by the graph neural network. The experiments with two datasets show better performance than other state-of-the-art works.

Also on computer vision, next article focuses on the problem of semantic interpretation of paintings. Some painters want to mean something with their work and they introduce signifiers that convey these meanings in their paintings. The research by Aslan and Steels focuses on the expression of meaning of paintings, exploring a comparative method to find the relationship between the source of a painting (e.g. a photography) and the painting itself. The authors investigated possible methods for aligning a painting and its source and used edge detection and the construction of comparative edge maps, to detect centers of interest. The article proposes a pipeline tested using paintings by the contemporary painter Luc Tuymans, focusing on showing the utility of the comparative method for semantic interpretation of a painting.

In the following article, we move from the area of computer vision to another fast developing area, which is natural language processing (NLP). In recent years, this area has experienced great advances due to large language models (LLM), and a new discipline called prompt engineering has emerged, whose purpose is to develop and optimize prompts for the efficient use of these LLM. Fine tuning by using the reinforcement learning from human feedback (RLHF) allows continuous improvement in the results obtained by the LLM. Pulari et al. propose a human selection strategy to improve the RLHF process in the news summarization problem. Multi-objective optimization is used for the trade-off between various objectives. Besides an evaluation metric H-Rouge (RH) is proposed for scenarios in which humans need to provide reviews and feedback. These human evaluations will facilitate an improved user experience, accurate summarizations, and reduced training costs.

The following articles correspond to a monograph section on the Effects of Culture on Open Science and Artificial Intelligence in Education, compiled and edited by Tlili, Burgos and Kinshuk. In the first article of this monograph, Tlili and Burgos discuss about the link between human and AI hallucinations and how its understanding could help to develop effective and safe AI systems to be used by everyone. Besides, they introduce the topic and the different articles of this monograph with insightful discussions on how cultural factors influence the adoption and implementation of open science and AI-driven technologies.

We hope that this issue fosters meaningful discussion and inspires further interdisciplinary research in these rapidly evolving fields.

Dr. Elena Verdú
Editor-in-Chief
Universidad Internacional de La Rioja

TABLE OF CONTENTS

OPEN ACCESS JOURNAL

ISSN: 1989-1660

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

COPYRIGHT NOTICE

A Smart Healthcare System Using Consumer Electronics and Federated Learning to Automatically Diagnose Diabetic Foot Ulcers

Sujit Kumar Das[1], Nageswara Rao Moparthi[2*], Suyel Namasudra[3], Rubén González Crespo[4*], David Taniar[5]

[1] Department of Computer Science and Engineering, Institute of Technical Education and Research, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, 751030, Odisha (India)
[2] Amrita School of Computing, Amrita Vishwa Vidyapeetham, Andhra Pradesh, Amaravati Campus (India)
[3] Department of Computer Science and Engineering, National Institute of Technology, Agartala, Tripura (India)
[4] Universidad International de La Rioja, Logroño (Spain)
[5] Faculty of Information Technology, Monash University (Australia)

* Corresponding author: mnrphd@gmail.com (N. R. Moparthi), ruben.gonzalez@unir.net (R. González Crespo).

Abstract

Privacy breaches on sensitive and widely distributed health data in consumer electronics (CE) demand novel strategies to protect privacy with correctness and proper operation maintenance. This work presents a scalable Federated Learning (FL) framework-based smart healthcare approach. Remote medical facilities frequently struggle with imbalanced datasets, including intermittent client connections to the FL global server. The proposed approach handled intermittent clients with diabetic foot ulcers (DFU) images. A data augmentation approach proposes to handle class imbalance problems during local model training. Also, a novel Convolutional Neural Network (CNN) architecture, ResKNet (K=4), is designed for client-side model training. The ResKNet is a sequence of distinctive residual blocks with 2D convolution, batch normalization, LeakyReLU activation, and skip connections (convolutional and identity). The proposed approach is evaluated for various client counts (5,10,15, and 20) and multiple test dataset sizes. The proposed framework can leverage consumer electronic devices and ensure secure data sharing among multiple sources. The potential of integrating the proposed approach with smartphones and wearable devices to provide highly secure data transmission is very high. The approach also helps medical institutions collaborate and develop a robust patient diagnostic model.

Keywords

I. Introduction

TECHNOLOGICAL advancements and globalization result in massive data collection by various enterprises and organizations using consumer electronic devices. The use of CE devices in data collection provides a great help in facilitating better service to humans. These data include a Wide range of information, from financial and industrial to medical records. However, in the case of medical records, the data transmissions from patients require more careful strategies. [1]. The greater demand for in-depth analysis of these vast data influx results in various advancements in machine learning (ML) and deep learning (DL) strategies [2], [3]. However, given the value of this information, ensuring the confidentiality and security [4] of the analyzed data is of utmost importance [5]. Adherence to regulatory requirements, such as the General Data Protection Regulation (GDPR)

[6], becomes mandatory in many instances. The traditional method for applying ML to decentralized data comprises a centralized framework sharing data by various entities, shown in Fig. 1. The client must transfer data to a centralized server for model training and subsequent results. Thereby, each client gets the final results. One of the major disadvantages of this approach is data confidentiality and the attributes related to it [7]. Furthermore, it requires a high bandwidth and low latency communication infrastructure to handle predictions promptly and successfully. One potential solution is each data owner possesses the model, so transferring data is not required when new information becomes accessible [8], [9]. It can decrease latency by making predictions for each client individually. Also, it reduces network reliance, lowering communication expenses. Nevertheless, each client must transmit data to the central server for the initial training of the model [10]. Although multiple traditional ML approaches are introduced to design

Fig. 1. Traditional centralized learning.

learning. The local machine update (LMU) is then sent to FL for global machine update (GMU). This collaborative approach helps better learning and provides a more realistic approach to ischaemia and infection identification in DFU with data privacy. The key contributions of the proposed method are as follows:

1. The use of FL to diagnose DFU without sharing sensitive data or related information in a remote healthcare setting to improve privacy and security.

2. A hybridized data augmentation technique is proposed to expand and balance the class distribution of the samples and examine the impact of the proposed augmentation approach on the collaborative framework.

3. A deeper residual block-based shallow CNN architecture that requires less computational resources for client-side model training.

4. The presence of decentralized data and the uneven distribution resulting from intermittent clients, the proposed architecture exhibits robustness and yields superior performance.

The remaining work sections are: Section II includes related works, and section III includes detailed problem definition and system model. Section IV offers a detailed description of the proposed methodology, outlining the key elements of the approach. Section V carries out experiments and provides detailed explanations, shedding light on the experimental process and results. Finally, section VI serves as the paper's conclusion, summarising the essential findings and insights gathered throughout the work.

II. Related Works

The development of DFU often leads to complications such as neuropathy and arterial disease in the lower extremities. It has been estimated that approximately 50%of DFU patients will experience neuropathy-related issues in later stages, with around 20% of them developing arterial blood flow problems. As many as 80% may suffer from both conditions simultaneously. However, identifying the presence of DFU solely based on its visual characteristics poses a significant challenge for clinicians. In many cases, DFU does not exhibit consistent shape and texture characteristics, making manual diagnosis unreliable [24], [25]. Manual diagnosis of DFU results in misdiagnosis in 2 out of every 3 cases. Therefore, it is imperative to harness advanced ML, DL, and computer vision techniques [26] to assist clinicians in accurately diagnosing DFU, thereby enhancing patient care. Developing an automatic diagnostic model can improve decision-making reliability at a minimal cost. While some research has been conducted in automatic DFU classification, there remains room for further investigation and development [27], [28].

Filipe et al. [16] introduced a two-stage ML classification approach that utilizes foot thermograms. At first, healthy and infected feet are distinguished [29]. In the next stage, assess the severity of the infection. However, the approach is costly and requires expertise to handle it. Goyal et al. [17] proposed DFUNet, a novel parallel convolution layer-based CNN architecture for differentiating between normal and abnormal DFU. DFUNet outperformed standard CNNs like LeNet, AlexNet, and GoogleNet, as well as traditional low-feature-based classification methods. However, the primary objective did not encompass identifying ischaemia in abnormal DFU wounds. Alzubaidi et al. [18] presented DFU_QUTNet, a CNN architecture for extracting multi-level features, which are subsequently input into Support Vector Machine (SVM) and K-Nearest Neighbours (KNN) classifiers. The method was compared to three standard CNN architectures (GoogleNet, AlexNet, and VGG16) to highlight its efficiency. Nonetheless, DFU_QUTNet did not address the identification of ischaemia in DFU cases. Another approach involved

data-driven DFU identification systems, most systems are centralized frameworks and work on imbalanced class labels. An approach has yet to consider a decentralized technique for model training. FL [11] gains popularity as a practical approach to guarantee that data remains on the servers of specific data owners, even throughout the training process [12]. It preserves essential data privacy when acquiring data locally from clients for centralized training is impossible [13]. FL does data analysis in a decentralized approach to avoid sending user data to central servers [14]. The primary motivation of this work is to address the disparity between advanced ML methods and real-world applicable conservative, large-scale healthcare solutions, taking better care of patient privacy and improving diagnostic performance in low-resource settings. To build such a robust system that can be put into practice on consumer devices for healthcare, the learning process has to be offloaded, and data sharing has to be internalized using federated learning. The development of DFU often leads to complications such as neuropathy and arterial disease in the lower extremities [15]. It is imperative to harness advanced ML, DL, and computer vision techniques to assist clinicians in accurately diagnosing DFU, thereby enhancing patient care. The identification of DFU using ML and DL techniques is introduced in various works. In one such work [16], a two-stage ML classification approach is proposed that analyzed foot thermograms. In another work [17], a novel parallel convolution layer-based CNN architecture (DFUNet) is proposed for differentiating between normal and abnormal DFU wounds. DFU_QUTNet [18], a CNN architecture is introduced for extracting multi-level features, which are subsequently input into Support Vector Machine (SVM) and K-Nearest Neighbours (KNN) classifiers. The hybridization of Neural Networks (NN) and Bayesian Classifiers (BNC) to detect necrotic tissue in wounds [19]. A dedicated DL-based model [20] for wound image segmentation with wound detection is also considered as a significant approach. Traditional ML approaches like a 2-level SVM classifier [21] for determining wound boundaries. In a study, standard CNN architectures [22] were employed for DFU wound segmentation. Later, a new dataset with ground truth labels for ischaemia and infection recognition [23].

This paper uses FL to investigate medical images and make predictions. While predictions are made, different numbers of clients and test image samples are considered. The proposed approach is decentralized by a training model with local augmented data. The local clients learn from the data they acquire and share the knowledge with other clients. At first, the imbalance data are augmented using a hybridized oversampling approach with combined capabilities of both SMOTE (Synthetic Minority Oversampling TEchnique) and SVM-SMOTE (Support Vector Machine-synthetic Minority Oversampling TEchnique). The proposed augmentation approach helps generate suitable synthetic data from the minority class and improves client

TABLE I. Gap Analysis of DFU Diagnosis Approaches

Study	Focus	Methods	Gaps
Filipe et al. [16]	DFU severity classification	Two-stage ML with foot thermograms	High cost; Expertise required
Goyal et al. [17]	DFU classification	Parallel convolution layer-based CNN (DFUNet)	Does not address ischaemia
Alzubaidi et al. [18]	DFU feature extraction	CNN + SVM and KNN classifiers	Does not address ischaemia
Veredas et al. [19]	Necrotic tissue detection	NN + Bayesian Classifiers (BNC)	High false positive rate; Efficiency concerns
Scebba et al. [20]	Wound image segmentation	DL-based model	Requires minimization of false positives
Wang et al. [21]	Wound boundary determination	2-level SVM classifier	Inefficiencies in level 1 classifier
Ohura et al. [22]	DFU wound segmentation	Standard CNN architectures	Does not address ischaemia detection
Goyal et al. [23]	Ischaemia and infection recognition	Traditional ML-based techniques and CNN	Less promising infection vs. non-infection results
Chen et al. [30]	Healthcare with Federated Learning	FL integration in healthcare	Not specific to DFU
Fathima et al. [31]	FL in healthcare	FL integration with IoT	No FL integration in DFU monitoring

the hybridization of Neural Networks (NN) and Bayesian Classifiers (BNC) to detect necrotic tissue in wounds [19]. The NN model extracted color and texture features from segmented wound images, which BNC then processed for prediction. This method requires strategies to reduce false positive detections and enhance efficiency. Scebba et al. [20] proposed a DL-based model for wound image segmentation, with wound detection performed before segmentation to improve generalization. However, similar to the previous approach, this method necessitates strategies to minimize false positive detections and enhance efficiency. Wang et al. [21] introduced a 2-level SVM classifier for determining wound boundaries. Incorrectly identified samples from level 1 undergo further processing by the level 2 SVM classifier to enhance overall performance. In a study by Ohura et al. [22], standard CNN architectures were employed for DFU wound segmentation, with U-Net achieving the best results among LinkNet, U-Net_VGG16, SegNet, and U-Net. Nevertheless, this approach also requires addressing ischaemia identification. Goyal et al. [23] introduced a new dataset with ground truth labels for ischaemia and infection recognition. They applied various traditional ML-based feature extraction techniques [24] and CNN architectures to differentiate between ischaemia and infection as binary classification problems [25]. An ensemble approach demonstrated significant performance improvements in both tasks, although infection vs. non-infection results were less promising compared to ischaemia vs. non-ischaemia classification. However, FL's use in DFU research has not yet been explored. But, in the healthcare domain, FL gained attention [30], [31]. Haya et al. [32] proposed frameworks integrating FL and the Internet of Things (IoT) within the healthcare domain. They introduced a data integration approach for monitoring patients remotely through IoT without incorporating FL into the surveillance process. The work is assessed using ECG data, validating that DL surpassed other implemented algorithms in performance. The work efficiently integrated FL with an IoT digital system to uphold personal privacy. Sun et al. [33] advocated using FL to enhance the learning efficiency of IoT-based intelligent automation. Numerous researchers [34] proposed specialized federated learning paradigms for detecting COVID-19 cases using X-ray images. They applied transfer learning on pre-trained algorithms, with residual networks exhibiting superior performance. Rahman et al. [35] introduced an FL model for healthcare that incorporates a DL edge layer and blockchain to enhance security and reliability. A system for sharing industrial IoT data using FL and blockchain is also proposed. In addition [36], proposed FL method for Electronic Health Records (EHRs) in the healthcare domain, showcasing promising results. Baheti et al. [37], leveraging FL, employed CT scans to detect respiratory lung nodules. Huang et al. [38] utilized a clustering technique to generate community-based data with clinical relevance, with their clustering-based FL model surpassing the standard FL model in performance. They addressed the issue of non-IID (Non-Independently and Identically Distributed) ICU health information by grouping clients into significant clinical populations, thus enhancing fatality and ICU wait-time predictions. Furthermore, Lee et al. [39] developed a system for patient resemblance learning within a federated environment while safeguarding patient privacy. Their model can identify similar patients across healthcare centres, even when no records are shared. The related works are summarised in table I.

III. Problem Definition and System Model

A. Problem Definition

The DFU are among the most serious diabetic complications and consequences often resulting in. These include limb shortening through amputation or surgery, lasting nerve pain, or severe infections. This makes early identification and diagnosis more imperative, but modern diagnostic systems have several limitations.

1. Data Privacy Concerns: A central requirement for deploying traditional diagnostic models is data storage and management at one central location, creating room for security risks concerning privately held information, especially with sensitive medical information.

2. Infrastructure Limitations: Many medical institutions may not have the necessary equipment to consolidate and analyse large amounts of data.

3. Limited access to expensive diagnostic tools: The current diagnostic systems are often associated with the need for specialized imaging modalities and do not offer many possibilities that could be available in every primary care and low-resource setting.

The objective of this work consists of designing an intelligent and decentralized DFU diagnosis supporting system overcoming the above-mentioned challenges due to the improvement of privacy, decreased need for centralized data, and broadening diagnostic tools' availability.

B. System Model

To deal with the abovementioned challenges, this work presents a novel system model for diagnosing DFU based on federated learning. The system allows each client's devices, such as mobile phones, tablets and wearables, to build local machine-learning models using local data without sharing raw data with the central facility. The key elements of the system model are defined as follows:

1. Client-Side Architecture:
   - Each client contains a local DFU image dataset of diabetic patients.
   - The ResKNet architecture is employed on the client side for model training. ResKNet is designed to be lightweight, efficient, and perform well in resource-constrained environments.
   - Clients use a hybridized data augmentation approach to handle the class imbalance in their datasets, generating additional synthetic data for underrepresented classes.

2. Federated Learning Framework:
   - The central server orchestrates the training process by collecting only the model updates (weights and gradients) from each client, instead of raw data.
   - The server aggregates these updates using techniques such as Federated Averaging (FedAvg) and applies them to the global model.
   - This decentralized approach ensures that patient data remains local, significantly enhancing privacy and reducing the risk of data breaches.

3. Data Transmission:
   - Client devices periodically transmit their locally trained model updates to the central server. These updates are secured using encryption protocols to ensure the confidentiality of sensitive medical data further.
   - The system is robust to intermittent client connections, common in resource-constrained environments.

4. Global Model Update:
   - Once the central server aggregates the model updates from multiple clients, the global model is updated and sent back to the clients for further training and improvement.
   - This iterative process continues until the global model achieves satisfactory accuracy for DFU diagnosis across all clients.

5. Consumer Electronics Integration:
   - The system is designed to be integrated with consumer electronics such as smartphones, wearable devices, and tablets, making it accessible to a wide range of users in various healthcare settings, including remote or under-resourced areas.
   - The reliance on readily available consumer electronics mitigates the need for expensive diagnostic tools, enabling scalable deployment.

IV. Proposed Method

The proposed method employs DFU images for infection detection and ischaemia identification. The motivation for the proposed scheme arises from the need to tackle several critical challenges in diagnosing DFU and healthcare diagnostics in general. These challenges revolve around privacy concerns, class imbalances in medical datasets, resource constraints in healthcare settings, and the high costs associated with traditional diagnostic tools. The proposed scheme is designed with these specific challenges in mind, leveraging advanced machine learning techniques in a way that is both efficient and scalable. The primary goal is to show how FL enables the secure and privacy-preserving sharing of crucial private information in CE devices when integrated alongside a CNN architecture. The proposed FL architectural system in CE, depicted in Fig. 2, is connected to remote hospitals through intermittent clients. Fig. 2 visualizes the overall flow and structure of the federated learning system. The left side of the figure illustrates the local training process on the client devices. Each device performs model training using locally collected data and applies data augmentation to handle class imbalances. The middle section of the figure shows the process of sharing model updates with the central server. Instead of raw data, only the updated model parameters are transmitted, ensuring privacy protection. The right side of the figure demonstrates the aggregation process performed by the central server. The server collects the updates from all participating clients, applies the FedAvg technique, and updates the global model. This global model is then redistributed back to the clients and used as the basis for further local training. This entire process iterates, gradually improving the global model's ability to diagnose DFU accurately while maintaining data privacy and minimizing the need for expensive infrastructure or specialized medical devices. These remote hospitals provide DFU images for the training procedure. Data augmentation guarantees data balance before application in local training machines to create LMU. Every hospital supplies LMU to the centralized server through local training weights. The centralized server, akin to a hub in a CE system, collects LMU from many remote hospitals and combines them to generate GMU. The GMU is then returned to the hospitals for updating to achieve precise classification findings. This collaborative exchange is reminiscent of the resonance in CE ecosystems, where information is shared for shared development. This work extensively tested the suggested method, considering scenarios involving intermittent clients and diverse image samples to determine its usefulness and make the best classification performance feasible. The proposed technique for decentralized model training in DFU images utilizing the FL approach is divided into five steps. These processes involve collecting datasets, augmentation, dealing with intermittent clients, client-side model training, and server-side model aggregation.

A. Dataset Preparation

The dataset is accessed from Manchester Metropolitan University, London. The creators compile the dataset from Lancashire Teaching Hospitals, London. The dataset is divided into two subdirectories: DFU images for detecting ischaemia (dataset 1) and infection (dataset 2) images. Both tasks rely on binary classification, which can aid in evaluating DFU wounds by identifying Ischaemia and Infection. The dataset 1 initially had 1459 complete foot pictures (210 Ischaemia and 1249 Non-ischaemia). Then, 1666 patches are extracted with the region of interest (ROIs) in mind. In dataset 2, the initial number of whole foot photos was 1459 (628 infected and 831 non-infected), and 1666 patches were created from them. A few example image patches from both datasets are shown in Fig. 3.

B. Data Augmentation

The data augmentation approach can efficiently address overfitting concerns and enhance the model's overall outcomes. Recent research has seen the emergence of various innovative approaches to advance the data augmentation field. Each customer might possess diverse image samples within each category in this situation. Consequently, this could lead to the potential problem of class imbalance. The dataset is comprised of 628 samples of infectious nature and 831 samples classified as normal. The normal class contains more samples, whereas the infectious class has a minor representation. The class imbalance problem is solved with hybridized oversampling techniques combining strategies of SMOTE and SVM-SMOTE. At first, the samples are divided equally. Secondly, in the first half of the samples, SMOTE oversampling was used to generate minority samples diagonally by choosing a random minority sample and its K-nearest neighbours. Thirdly, the SVM-SMOTE oversampling method was applied to half of the samples left out. At last, separately generated synthetic samples are combined to get a balanced and more representative and balanced dataset. The data generation has been conducted in the following steps:
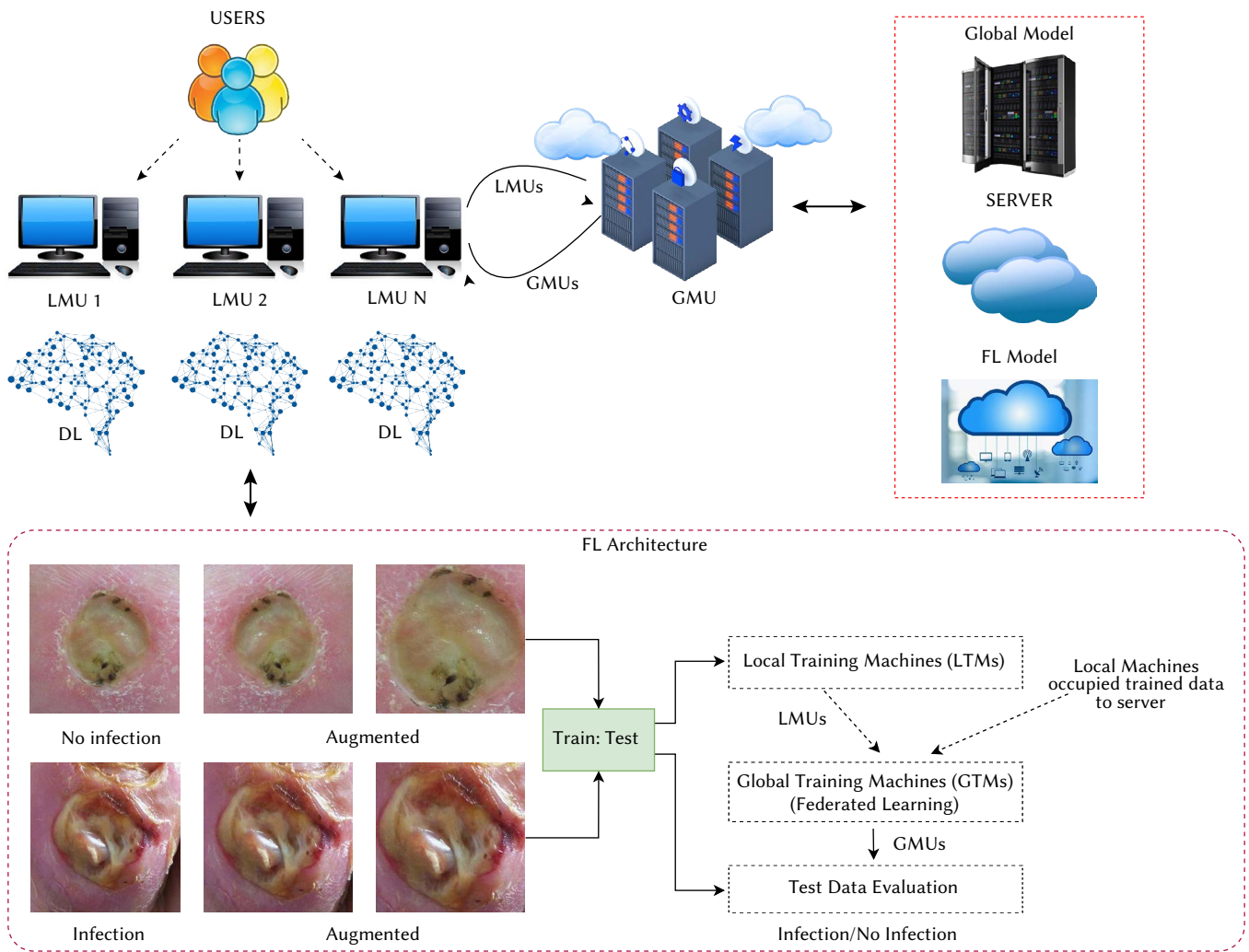
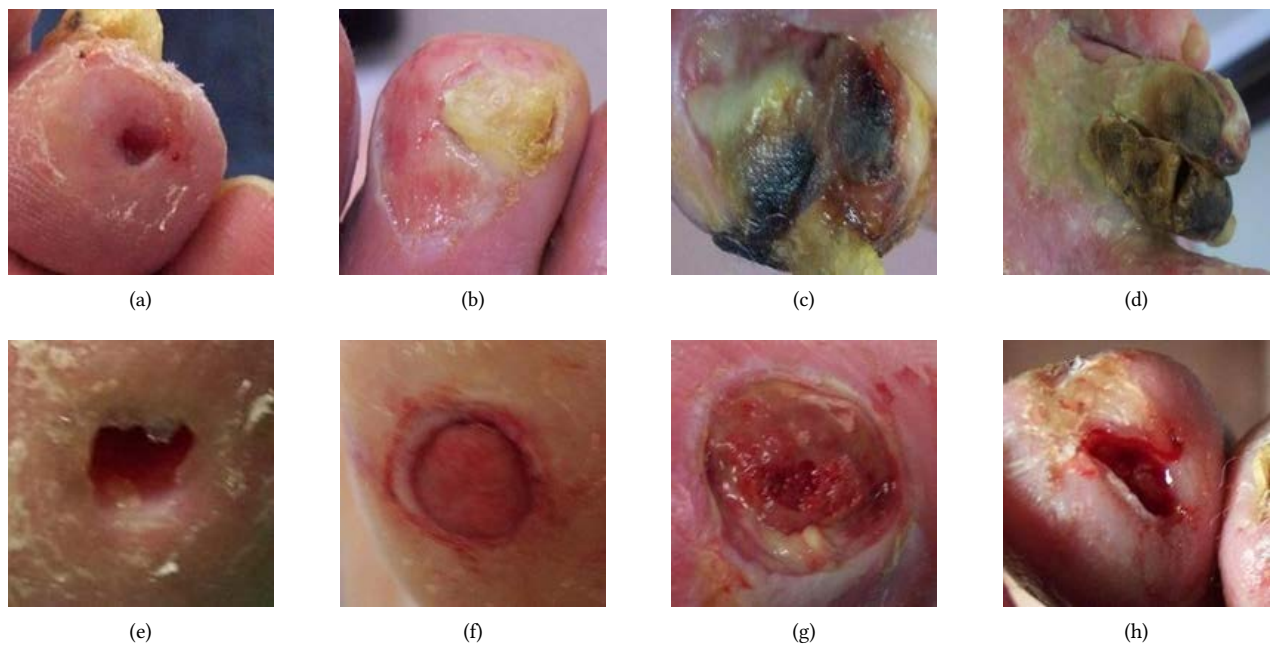Fig. 2. Overview of the federated learning framework for DFU diagnosis.

Fig. 3. Example images: (a)-(b) Non-Ischaemia, and (c)-(d) Ischaemia images from DFU Ischaemia (dataset 1). (e)-(f) Non-Infection, and (g)-(h) Infection images from DFU Infection (dataset 2).

1. Divide the training samples T into T_1 and T_2, where $|T_1| = |T_2|$.

2. For each sample in T_1, do the following:

   - Choose any minority sample *'r'* from the feature space.
   - Select any instance *'m'* from the k nearest neighbors of *'r'*.
   - Generate a synthetic sample S_{SMOTE} along the line segment between *'r'* and *'m'*, where $S_{SMOTE} = r + w \times (m - r)$, and w is a random number in the range [0, 1].

3. For each sample in T_2, do the following:

   - Train an SVM with T_2 to find the decision boundary.
   - Choose a minority sample *'r'*.
   - Select any instance *'m'* from the k nearest neighbours of *'r'* within the decision boundary line segment.
   - If the number of majority samples among the k nearest neighbors is less than half:
     - Generate a synthetic sample $S_{SVMSMOTE}$ either above or below the line segment connecting *'m'* and *'r'* (extrapolation).
   - Else, generate a synthetic sample *'s'* between the line segment connecting *'m'* and *'r'* (interpolation).

4. Combine the synthetic samples generated in steps 2 and 3 to create a new set of training samples T'.

C. Irregular Clients

Several variables can contribute to the issue of infrequent clients. The most prevalent issues typically revolve around constraints in data transmission, network connections, and computing infrastructure [25]. Here, approaches are employed to address irregular clients, and the imbalance dataset is used.

- The proposed method is being tested on various clients, with some departing and others joining. The weights obtained during training sessions are not considered if a client departs from the system. The weights from the new client are integrated into the aggregation. In this scenario, the model performance is influenced by the image samples provided by the new client. This approach can yield improved classification results if the new client possesses sufficient picture samples for local training. However, it is important to note that new clients with fewer picture samples may negatively impact the aggregation weights.

- Upon a client's departure, its latest weights are retained and used in subsequent aggregations to update the model. The weights obtained from the most recent client are added to the aggregation, and the departing client's image samples are used to evaluate the model's performance.

D. CNN for Client-Side Model Training

On the client side, each FL client utilized personal data and local resources to execute mini-batch ADAM and local CNN training. The algorithm 1 is used for local client training. In algorithm 1, the inputs $Weight$ refer to the local model weight, and $Weight_T$ refers to the global model weight at round t. D_b is the data size in batches and DP_k is data points on client K. The local data trains the local model with collected weight $Weight_T$. Once the weights are collected, the updated $Weight_T$ the $Weight$ is updated. After iteratively running ADAM with local epochs aligned to create the most recent model update, the client computes a gradient update. The newly updated parameters are subsequently transmitted to the global server to update the data stored on the server. Further, the significant role played by the proposed CNN. The proposed CNN architecture, Res4Net, is designed as a shallow network with a deeper structure based on residual blocks. This network comprises a sequence of distinctive residual blocks involving 2D convolution, batch normalization, and LeakyReLU activation, connected by skip connections (convolutional and identity). A visual representation of the model's layer-by-layer architecture can be observed in Fig. 4. The first residual block output block with skip connection can be defined mathematically from eq. (1).

$$Res4Net_{block(1/3)} = ADD[((Convskip_1^{W \times H \times D}, BN),$$
$$(Conv2D_1^{W \times H \times D}, BN, LR), (Conv2D_3^{W \times H \times D}, BN, LR),$$
$$(Conv2D_1^{W \times H \times D}, BN))]LR \tag{1}$$

The next consecutive residual block contains no skip connection and can be derived mathematically by eq. (2).

$$Res4Net_{block(2/4)} = ADD[((Conv2D_1^{W \times H \times D}, BN, LR),$$
$$(Conv2D_3^{W \times H \times D}, BN, LR), (Conv2D_1^{W \times H \times D}, BN))]LR \tag{2}$$

where $Res4Net_{block(1/3)}$ is the output of residual blocks 1 and 3. ADD represents addition operation $Convskip\_1^{W \times H \times D}$ is skip convolution layer with width W, height H, and D channel depth. The BN stands for batch normalization, and LR stands for leakyReLU activation function. After the convolution operation, The batch normalisation output will help balance input feature map distribution. The output of BN operation $Out_{B,C,X,Y}$ is derived in eq. (3).

$$Out_{B,C,X,Y} = \gamma_C \frac{Input_{B,C,X,Y} - \mu_C}{\sqrt{\sigma_C^2 + \epsilon}} + \beta_C \ \forall B, C, X, Y \tag{3}$$

where $Input_{B,C,X,Y}$ is a four-dimensional input with batch (B), Channel (C), X, and Y are spatial dimensions. The μ_C represents mean activation and β_C and σ_C are channel-wise affine transmission.

Algorithm 1: Client-Side Model Training (LMU)

Input: (**Weight**, **Weight**$_T$).

Output: Local Model Update (LMU) Weight

1:   **Begin (Weight = Weight**$_T$)   // **Initialization**

2:   **Spit** $D_b \leftarrow DP_K$, where D_b is batch data size and DP_K is data points for client K.

3:   **Update Weight**$^{(T, D)}$ with ADAM optimizer and initial learning rate $1e - 2$.

4:   **for** local epochs i from 1 to N: **do**   // **Beginning of outer for loop**

5:     **while** use optimizer: **do**   // **Optimizer loop**

6:       **for** every $D_{(bi)}$ in D_i: **do**   // **Data batch loop**

7:         **Find** $Global^D_{(bi)} \leftarrow \sigma D(Weight_b, N)$.   // **Global update**

8:         **Save Weight** $\leftarrow$ **Weight**$_b \leftarrow Global^D_{(bi)}$.   / **Weight update**

9:       **end for**   // **End of data batch loop**

10:    **end while**   // **End of optimizer loop**

11: **end for**   // **End of outer for loop**

12: **return Weight** as LMU.   // **Return LMU**

E. FL Server-Side Model Aggregation

The LMU is the weight after training clients with irregular clients. The LMU from the client side forms GMU on the FL server side. The details of GMU formation are discussed in algorithm 2. If C_N is the client's number and C_k is the number of client data for k, $k \ \forall \ [1, ..., C_N]$, then the average weight from clients is calculated with eq. (4).

$$W_{avg(k)} = \frac{C_1}{\sum_{k=1}^{C_N} C_k} \tag{4}$$

The aggregated weights on the FL server side can be calculated using eq. (5).

$$W_{aggregated} = \sum_{k=1}^{C_N} C_k W_k \tag{5}$$
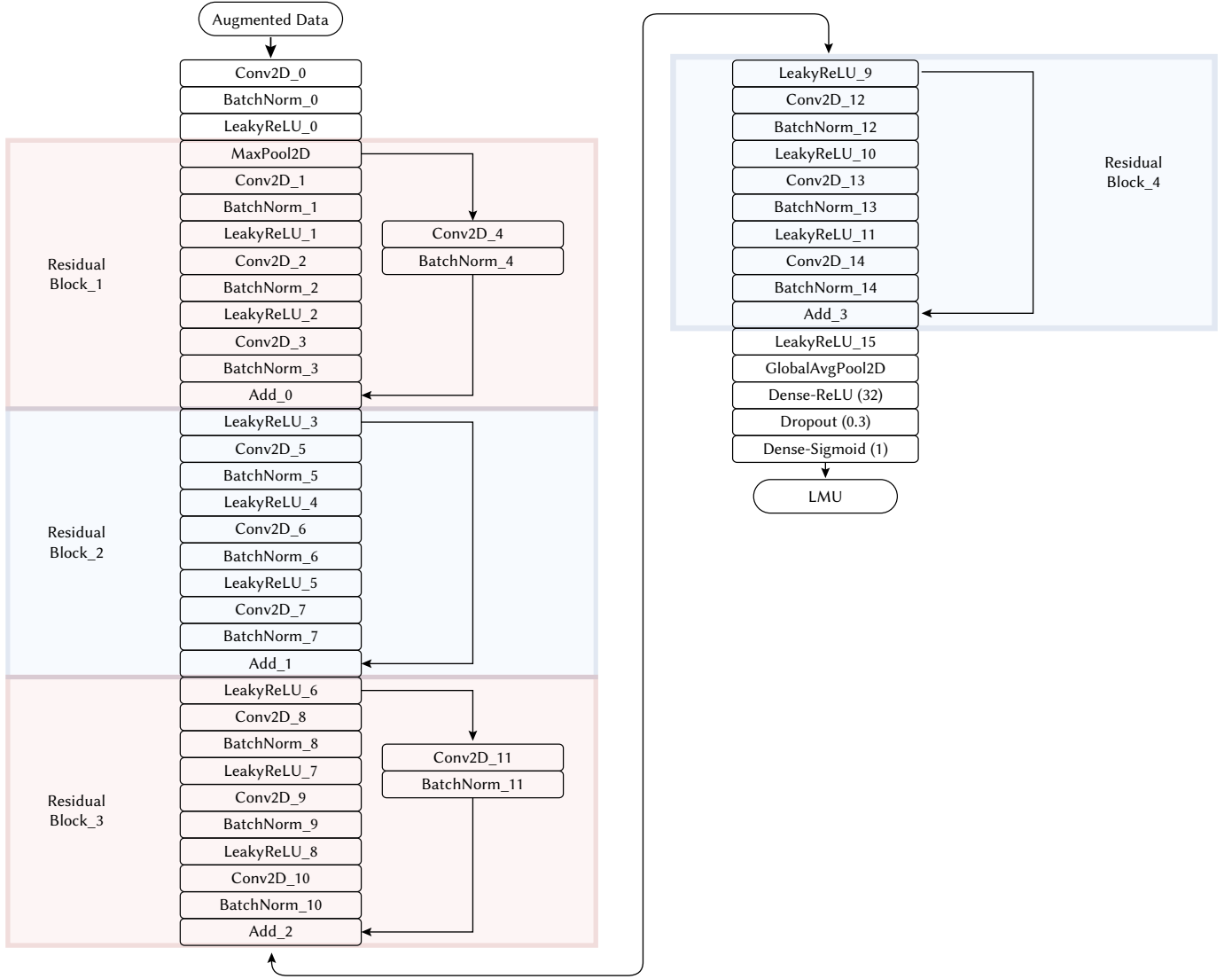
Fig. 4. The proposed CNN (ResKNet) architecture for client training.

Algorithm 2: FL Server-Side Aggregation (GMU)

Input: **Weight**, FL_n (Federated cycles).

Output: **Weight**$_{agg}$, weight aggregation.

1: **Begin** $FL_n \leftarrow 0$, **Weight**^{0}, both cycles and weight initialized to 0.

2: **Compute** $M \leftarrow MAX(C \times k, 1)$ maximum clients.

3: **Select** I_t randomly of n clients at t cycle.

4: **while** $k \leftarrow I_t$: **do**   // **Beginning of while loop**

5:    Update weight **Weight**$_{t-1}$ to I_{t-1}.

6:    **Weight**$^t_{tk} \leftarrow update(t^k, \text{Weight}_{t-1})$.   // **Update current weight**

7: **end while**   // **End of while loop**

8: **Aggregation Weight**$_{agg} \leftarrow \sum_{t=1}^{t^k} \frac{n_t}{n}$**Weight**$^t_{tk}$.   // **Aggregating weights**

9: **return Weight**$_{agg}$

V. Performance Analysis

The proposed model performance is evaluated with the help of multiple important evaluation metrics. Further, the results are represented and analyzed with the help of various tables and graphs.

The following subsections include a detailed discussion of results evaluation and discussion.

A. Evaluation Metrics

The proposed approach is evaluated for varying-sized test data from intermittent clients. The proposed approach tested the capability of identification of infection vs. non-infection DFU wounds. Five evaluation metrics are recorded to check the performance of the proposed approach. The evaluation matrices considered Accuracy, Precision, Sensitivity, Specificity, and F1-Score are given in eqs. (6) - (10).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Sensitivity/Recall = \frac{TP}{TP + FN} \quad (8)$$

$$Specificity = \frac{TN}{TN + FP} \quad (9)$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (10)$$
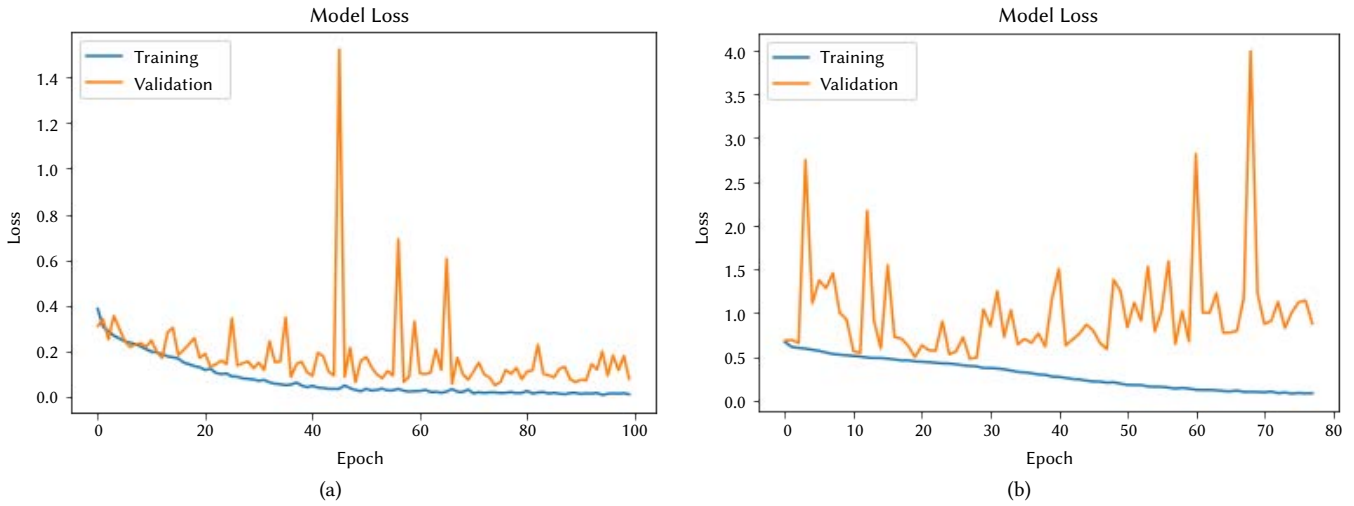
Fig. 5. Training and validation loss curve: (a) DFU Ischaemia (dataset 1). (b) DFU Infection (dataset 2).

TABLE II. Results With Different Clients Using Dataset 1

Class	Client	Precision	Sensitivity	Specificity	Accuracy	F1-Score
Ischaemia	5	96.44	97.43	96.46	96.90	97.47
Non-Ischaemia		96.74	96.84	96.97		
Ischaemia	10	97.15	97.20	97.18	97.25	97.55
Non-Ischaemia		97.06	97	97.07		
Ischaemia	15	98.37	98.67	98.38	98.73	98.41
Non-Ischaemia		97.87	98.27	97.88		
Ischaemia	20	99.18	99.49	99.20	99.50	99.36
Non-Ischaemia		98.98	99.18	99.00		

In eqs. (6)-(10), *TP* refers to True Positive, *TN* refers to True Negative, *FP* refers to False Positive, and *FN* refers to False Negative counts.

B. Results and Discussion

The data collected from consumer electronic devices to train an FL framework to learn from data can be analysed using various parameters. One such measure is the training vs. validation loss curve. The training vs. validation loss curves on the server side are shown in Fig. 5 for both datasets. Fig. 5 (a) is the loss curve for dataset 1 (ischaemia vs non-ischaemia), which shows training loss in blue coloured and validation loss in orange coloured. The training loss constantly decreases, but the validation loss shows a more dynamic nature. On the other hand, in the case of dataset 2, a more suitable training and validation loss curve is achieved. In Fig. 5 (b), for dataset 2, the training loss linearly decreases. However, in the case of validation loss, a sharp increase is observed in the 45th epoch. Additionally, the proposed method examines intermittent clients, selecting a certain amount of clients in every run while utilizing 300 instances from both datasets as the test data. Table II reports the classification results of dataset 1 for 5, 10,15, and 20 clients. In case 5 clients' classification between ischaemia and non-ischaemia, the precision, sensitivity, and specificity scores are 96.44%, 97.43%, and 96.46%, respectively. However, for non-ischaemia classes, a small improvement of the results is observed between 1-2% for 5 clients. Furthermore, the accuracy and F1-score values are the same for both classes at 96.90% and 97.47%, respectively. Similarly, for clients, 10 table II indicates that on increasing the number of clients to 10, classification performance increases. The classification results are improved (1-2)% compared to 5 clients. The results suggest that the sensitivity score of the non-ischaemia class is slightly higher than the ischaemia class. This is the reason for the larger value of false negative counts. However, the interesting observation from table II is that an increase in client numbers increases the performance of the proposed system. The reason behind the performance improvement is an increase in client numbers reduces the load. Further, with 15 clients for dataset 1, the highest score in the ischaemia class is 98.37%, 98.67%, and 98.38% for precision, Sensitivity, and Specificity, respectively. These results are again better than 5 and 10 clients. Another important observation is that the proposed approach achieved higher performance in ischaemia identification than non-ischaemia identification for 15 clients. The last setup results with client 20 are reported in table II. The proposed approach achieved almost perfect results in identifying both ischaemia and non-ischaemia. In the case of the ischaemia class, the highest precision, recall, and specificity are 99.18%, 99.49%, and 99.20%, respectively. Similarly, with a small low score for the non-ischaemia class, the highest precision, sensitivity, and specificity scores are 98.99%, 99.18%, and 99.00%, respectively. The overall accuracy score is 99.50%, with an F1-Score of 98.41. Therefore, starting from 5 clients to 20 clients, the results are improved, which signifies that an increase in clients helps in better learning and thereby improves performance.

Result table III shows the performance of the proposed approach with 5, 10, 15, and 20 clients in dataset 2. The classification results of infection are poor compared to the ischaemia classification. The highest results for the infection class in terms of precision, recall, and specificity are 85.90%, 82.81%, and 85.36%, respectively. Similarly, table III shows results with 10 clients for dataset 2. In the case of dataset 2 for infection and non-infection, both classes show a sharp increase in results. The results are improved by more than 3%. However, in the case of 10 clients, the identification results of the non-infection class slightly decreased.

TABLE III. Results With Different Clients Using Dataset 2

Class	Client	Precision	Sensitivity	Specificity	Accuracy	F1-Score
Infection	5	85.90	82.81	85.36	84.04	84.86
Non-Infection		86.80	83.46	86.42		
Infection	10	88.59	83.73	88.18	87.09	85.33
Non-Infection		88.03	84.28	87.65		
Infection	15	92.84	89.19	92.59	91.34	90.32
Non-Infection		91.92	87.56	91.71		
Infection	20	99.95	94.10	97.02	96.26	94.49
Non-Infection		95.55	91.65	96.47		
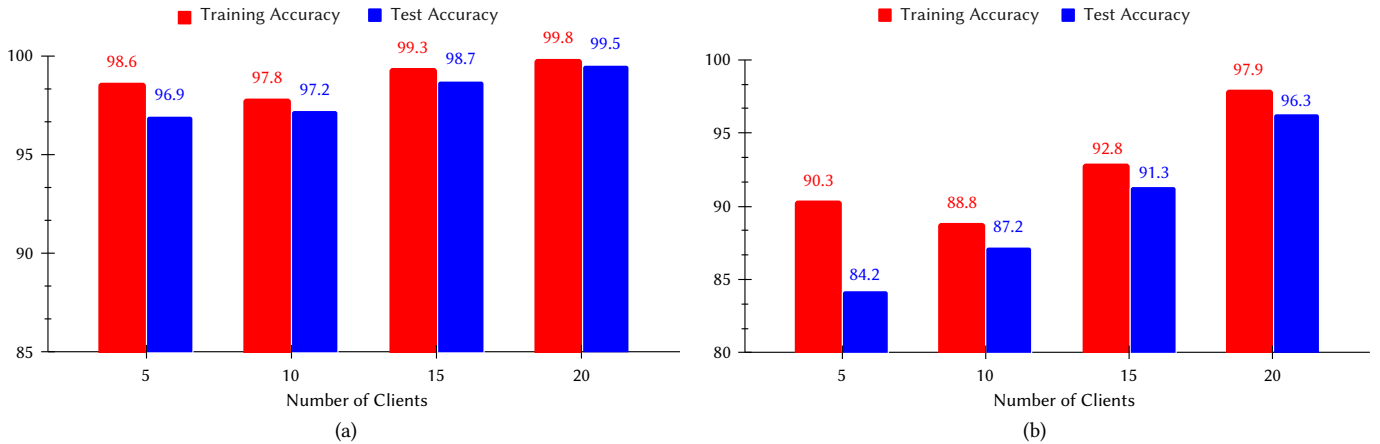
Fig. 6. The training and test accuracy on different numbers of clients: (a) DFU Ischaemia (dataset 1). (b) DFU Infection (dataset 2).

Once the number of clients further increased to 15, the performance of the proposed approach improved, as shown in table III. But the increment is quite promising, with an average value of (4-5)%. In the case of infection identification, the results are reported as 92.84%, 89.19%, and 92.59% for precision, sensitivity and specificity. The improvement in sensitivity score is lower than the other two evaluation metrics due to the large count of false negative values. Similarly, an improvement is observed in the case of non-infection identification compared to 10 clients. However, the improvement is slightly lower than the infection class. The identification of positive class always has a higher priority in medical diagnosis.

The last records in table III show the results of dataset 2 with 20 clients. There is a significant improvement in both infection and non-infection identification. The results with 20 clients are (4-7)% higher than 15 clients. Also, the results are improved to around 4% for the non-infection class. The increase in clients helps improve the results of the proposed approach. However, in the case of dataset 2, where identification of the correct class is more complex than dataset 1, there are more improvements. Further, the training and test accuracy in the line graph for different client numbers (dataset 1) is shown in Fig. 6 (a). The difference between training accuracy and test accuracy reduces the client number is increased. Similarly, Fig. 6 (b) shows the train and test accuracy for different clients in dataset 2. The characteristic of the line graph is similar to dataset 1. The reduction in train and test accuracy differences from 10 clients shows that the model learns very well and provides more generalization once the client numbers are increased.

In FL architecture, collaborating hospitals can produce LMUs using a variety of image sets as test data. Therefore, evaluating the proposed approach with random test data is very important. The proposed model is analyzed using varied test data sizes from both datasets. The results for dataset 1 on taking various image sample sizes are reported in table IV. The different numbers of image test samples are taken as 200, 150, 100, and 50. In the case of 200 test samples, the overall accuracy is 98.22%, and the F1-score is 98.91%. The precision, sensitivity and specificity scores in the ischaemia class are slightly higher than the non-ischaemia class. Further, when the test sample size is reduced to 150, the overall accuracy and F1-score scores are reduced to 97.46% and 97.05%, respectively. Similar characteristics are observed in the case of ischaemia and non-ischaemia, where precision, sensitivity and precision are reduced by around 1%. The highest overall accuracy and F1 scores are achieved for a test sample size of 100 with the values of 99.54% and 99.12%, respectively. With 100 test sample size, the performance of ischemia identification is better than non-ischaemia identification. In the case of ischaemia identification, the highest sensitivity score is 99.08%. Similarly, the scores of precision and specificity are 99.18% and 99.10%, respectively. The performance of non-ischaemia with a 100 test sample size is slightly reduced, but it is the highest result among other test sample sizes. The precision and sensitivity scores for non-ischaemia are 98.77%. The specificity score is also almost the same, with a value of 98.80%. Further, when the test sample is reduced to 50, the overall accuracy and F1-score outperformed compared to the 200 and 150 sample sizes. More specifically, with an accuracy of 99.03%, it is the second-best performing sample size. The sensitivity score of the ischaemia class is 98.67%, whereas for non-ischaemia, it is 98.27%. The results of considered test samples for dataset 2 are reported in table V. Similar characteristics are observed for infection vs. non-infection datasets as well. The overall accuracy and F1-score of 200 and 150 test samples are poor compared to 100 and 50 sample sizes. In the case of 200 test samples, the accuracy score is achieved as 85.86%, Which is further reduced to 83.71% with a 150 sample size. The F1-score for 200 samples is reported as 85.52%, and the lowest F1-score of 81.05% is reported with a 150 sample size. In individual classes, the precision, recall, and specificity scores for infection are 86.58%, 83.46%, and 86.06%,

TABLE IV. Results With Different Test Data Size Using Dataset 1

Class	Client	Precision	Sensitivity	Specificity	Accuracy	F1-Score
Ischaemia	200	97.67	98.16	97.69	98.22	98.91
Non-Ischaemia		97.36	97.86	97.39		
Ischaemia	150	96.67	97.45	96.69	97.46	97.05
Non-Ischaemia		96.16	96.84	96.19		
Ischaemia	100	99.08	99.18	99.10	99.54	99.12
Non-Ischaemia		98.77	98.77	98.80		
Ischaemia	50	98.57	98.67	98.60	99.03	98.61
Non-Ischaemia		98.27	98.27	98.30		

TABLE V. Results With Different Test Data Size Using Dataset 2

| Class	Client	Precision	Sensitivity	Specificity	Accuracy	F1-Score
Infection	200	86.58	83.46	86.06	85.86	85.52
Non-Infection		87.24	83.96	86.06		
Infection	150	84.51	82.16	83.77	83.71	81.05
Non-Infection		85.13	82.48	83.06		
Infection	100	97.10	93.28	97.00	96.31	94.51
Non-Infection		96.25	92.47	96.11		
Infection	50	90.83	87.56	90.31	90.38	88.01
Non-Infection		89.98	86.74	89.59		
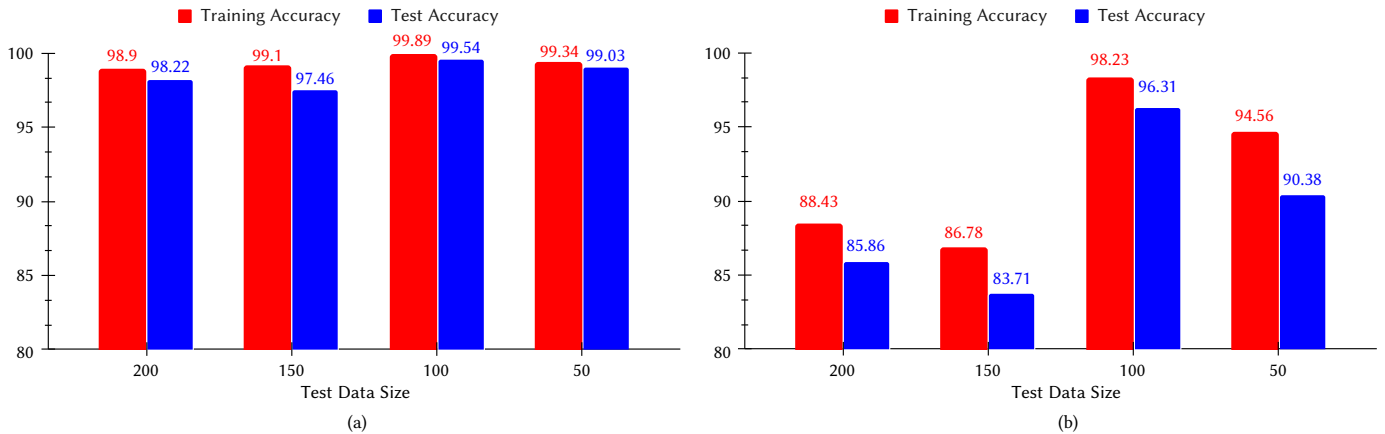
Fig. 7. The training and test accuracy on different test data sizes: (a) DFU Ischaemia (dataset 1). (b) DFU Infection (dataset 2).

respectively, with 200 samples. These scores slightly improved for the non-ischaemia class with around 1%. Among all considered sample size settings, the highest results are reported with a 100-test sample size. The overall accuracy and F1-score are 96.31% and 94.51%, respectively. In the case of infection with 100 samples, the highest precision, sensitivity and specificity results are 97.10%, 93.28%, and 97.00%. A slightly low sensitivity score compared to precision and specificity is due to a somewhat high value of false negative count. Similarly, in the case of non-infection identification, the results are impressive, with a slight decrement around (1-2)% compared to infection. The second-best result is achieved with a sample size of 50, where the accuracy and F1-Score are 90.38% and 88.01%. The evaluation of different test data sizes for dataset 1 and dataset 2 is shown in Figs. 7 (a) and 7 (b), respectively. The difference between training and test accuracy for 200 and 150 is higher than 100 and 50 in both datasets. Therefore, the observation from these tables is that the proposed approach can greatly help evaluate small sample sizes in real-life clinical practice. Fig. 8 shows the comparison of the proposed approach in terms of

accuracy and F1-Scores with state-of-the-art works. The proposed approach outperforms the popular standard CNNs. The significant improvements of the proposed work with nearly (4-9)% of accuracy in dataset 1 and (2-24)% in dataset 2 shows its importance in diagnosing the disease.

VI. Conclusions

This paper presents a federated learning-based approach for the automatic diagnosis of DFU, addressing key challenges such as data privacy and diagnostic accuracy. By leveraging a decentralized learning framework, this work enables training machine learning models directly on client devices, such as smartphones and tablets, without transferring sensitive patient data to a central server. This approach significantly mitigates privacy concerns commonly associated with centralized data processing in healthcare. Furthermore, the proposed approach introduces a hybridized data augmentation technique to handle class imbalance in DFU datasets, improving the model's ability
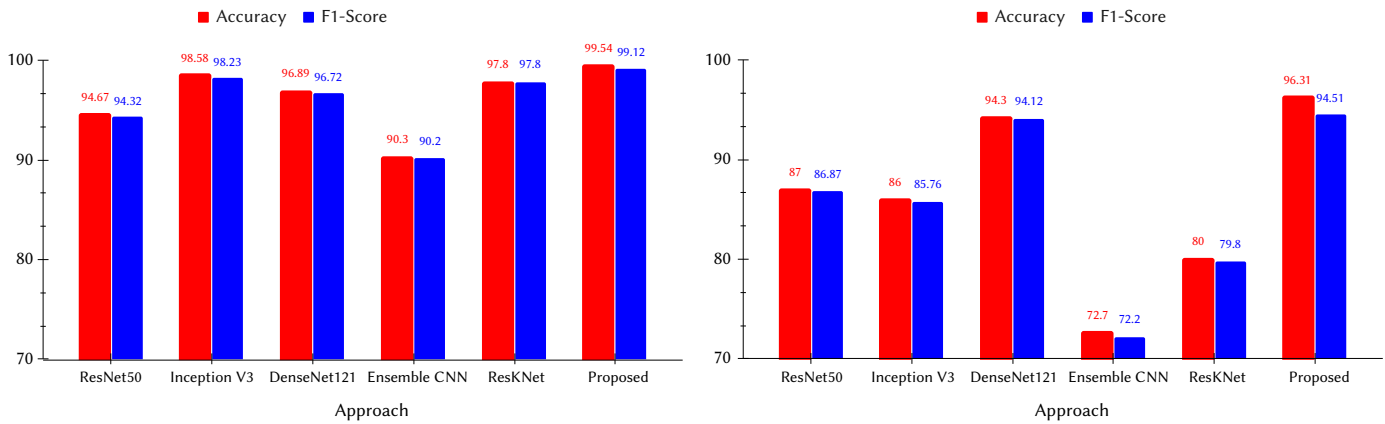
Fig. 8. Comparison with SOTA works: (a) DFU Ischaemia (dataset 1). (b) DFU Infection (dataset 2).

to classify ischaemia and infection in DFU. Using lightweight CNN (ResKNet) demonstrates the feasibility of running effective models even on resource-constrained devices, offering a practical solution for real-world healthcare applications. The result shows that the federated learning system achieved strong performance across multiple communication rounds, with continuous improvements in accuracy and reductions in model loss. The system's effectiveness is further validated with high precision in correctly classifying DFU stages, with minimal false positives and false negatives. This performance, coupled with the system's scalability and ability to function in low-resource environments, underlines its potential for widespread deployment in developed and developing regions. Ultimately, the proposed approach has the potential to provide accessible, affordable, and privacy-preserving diagnostic support for DFU patients. Future research could expand on this work by exploring more sophisticated data augmentation techniques to further enhance model performance, particularly in scenarios with more severe class imbalances. Applying the federated learning framework to other medical domains may improve data privacy and diagnostic accuracy across various conditions. These developments could help strengthen the impact of federated learning in healthcare, making it a key enabler for secure and effective medical diagnostics in diverse settings.

VII. Declaration of Competing Interest

The authors declare that they have no conflict of interest.

VIII. Authors Contribution Statement

Sujit Kumar Das is the main author of this paper, who has conceived the idea and discussed it with all co-authors. He has developed all the algorithms. Nageswara Rao Moparthi and Suyel Namasudra have performed the simulations of this paper. Rubén González Crespo is the corresponding author, who has supervised the entire work and proofread the paper. David Taniar has evaluated the performance and write-up of this work.

IX. Compliance With Ethical Standards

The authors did not use animals and human participants in the study reported in this work.

X. Data Availability

The data that support the findings of this study are available with the authors, but restrictions apply to the availability of these data.

Thus, data are not publicly available. However, data are available with permission from the Department of Computing and Mathematics, Manchester Metropolitan University, UK.

XI. Funding

References

[1] A. K. Mishra, P. Roy, S. Bandyopadhyay, S. K. Das, "Feature fusion based machine learning pipeline to improve breast cancer prediction," *Multimedia Tools and Applications*, vol. 81, no. 26, pp. 37627–37655, 2022, doi: 10.1007/s11042-022-13498-4.

[2] S. Saminu, G. Xu, S. Zhang, I. Ab El Kader, H. A. Aliyu, A. H. Jabire, Y. K. Ahmed, M. J. Adamu, "Applications of artificial intelligence in automatic detection of epileptic seizures using eeg signals: A review," in *Artificial Intelligence and Applications*, vol. 1, 2023, pp. 11–25.

[3] S. Jiang, Y. Gu, E. Kumar, "Magnetic resonance imaging (mri) brain tumor image classification based on five machine learning algorithms," *Cloud Computing and Data Science*, vol. 4, pp. 122–133, 2023, doi: 10.37256/ccds.4220232740.

[4] R. J. Robinson, "Insights on cloud security management," *Cloud Computing and Data Science*, vol. 4, pp. 212–222, 2023, doi: 10.37256/ccds.4220233292.

[5] J. Purohit, R. Dave, "Leveraging deep learning techniques to obtain efficacious segmentation results," *Archives of Advanced Engineering Science*, vol. 1, no. 1, pp. 11–26, 2023, doi: 10.47852/bonviewAAES32021220.

[6] P. Voigt, A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, no. 3152676, pp. 10–5555, 2017, doi: 10.1007/978-3-319-57959-7.

[7] S. Namasudra, P. Roy, "Size based access control model in cloud computing," in *Proc. of the International Conference on Electrical, Electronics, Signals, Communication and Optimization*, 2015, pp. 1–4.

[8] A.-R. Al-Ali, I. A. Zualkernan, M. Rashid, R. Gupta, M. AliKarar, "A smart home energy management system using iot and big data analytics approach," *IEEE Transactions on Consumer Electronics*, vol. 63, no. 4, pp. 426–434, 2017, doi: 10.1109/TCE.2017.015014.

[9] S. H. Requena, J. M. G. Nieto, A. Popov, I. N. Delgado, "Human activity recognition from sensorised patient's data in healthcare: A streaming deep learning-based approach," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 1, pp. 23–37, 2023, doi: https://hdl.handle.net/10630/27669.

[10] Y. He, X. Jin, Q. Jiang, Z. Cheng, P. Wang, W. Zhou, "Lkat-gan: A gan for thermal infrared image colorization based on large kernel and attentionunet-transformer," *IEEE Transactions on Consumer Electronics*, vol. 69, no. 3, pp. 478 – 489, 2023, doi: 10.1109/TCE.2023.3280165.

[11] F. Ullah, G. Srivastava, H. Xiao, S. Ullah, J. C.- Lin, Y. Zhao, "A scalable federated learning approach for collaborative smart healthcare systems with intermittent clients using medical imaging," *IEEE Journal of*

Biomedical and Health Informatics, 2023, doi: 10.1109/JBHI.2023.3282955.

[12] Z. Zhang, Y. Li, Y. Gong, Y. Yang, S. Ma, X. Guo, S. Ercisli, "Dataset and baselines for IID and OOD image classification considering data quality and evolving environments," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 1, pp. 6-12, 2023, doi: 10.9781/ijimai.2023.01.007.

[13] H. Guan, P.-T. Yap, A. Bozoki, M. Liu, "Federated learning for medical image analysis: A survey," *Pattern Recognition*, p. 110424, 2024, doi: 10.1016/j.patcog.2024.110424.

[14] V. Stephanie, I. Khalil, M. Atiquzzaman, X. Yi, "Trustworthy privacy-preserving hierarchical ensemble and federated learning in healthcare 4.0 with blockchain," *IEEE Transactions on Industrial Informatics*, vol. 19, pp. 7936 – 7945, 2022, doi: 10.1109/TII.2022.3214998.

[15] S. K. Das, S. Namasudra, A. K. Sangaiah, "Hcnnet: hybrid convolution neural network for automatic identification of ischaemia in diabetic foot ulcer wounds," *Multimedia Systems*, vol. 30, no. 1, p. 36, 2024, doi: 10.1007/s00530-023-01241-4.

[16] V. Filipe, P. Teixeira, A. Teixeira, "Automatic classification of foot thermograms using machine learning techniques," *Algorithms*, vol. 15, no. 7, p. 236, 2022, doi: 10.3390/a15070236.

[17] M. Goyal, N. D. Reeves, A. K. Davison, S. Rajbhandari, J. Spragg, M. H. Yap, "Dfunet: Convolutional neural networks for diabetic foot ulcer classification," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 5, pp. 728–739, 2018, doi: 10.1109/TETCI.2018.2866254.

[18] L. Alzubaidi, M. A. Fadhel, S. R. Oleiwi, O. Al- Shamma, J. Zhang, "Dfu_qutnet: diabetic foot ulcer classification using novel deep convolutional neural network," *Multimedia Tools and Applications*, vol. 79, no. 21, pp. 15655–15677, 2020, doi: 10.1007/s11042-019- 07820-w.

[19] F. Veredas, H. Mesa, L. Morente, "Binary tissue classification on wound images with neural networks and bayesian classifiers," *IEEE transactions on medical imaging*, vol. 29, no. 2, pp. 410–427, 2009, doi: 10.1109/TMI.2009.2033595.

[20] G. Scebba, J. Zhang, S. Catanzaro, C. Mihai, O. Distler, M. Berli, W. Karlen, "Detect-and-segment: a deep learning approach to automate wound image segmentation," *Informatics in Medicine Unlocked*, vol. 29, p. 100884, 2022, doi: 10.1016/j.imu.2022.100884.

[21] L. Wang, P. C. Pedersen, E. Agu, D. M. Strong, B. Tulu, "Area determination of diabetic foot ulcer images using a cascaded two-stage svm-based classification," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2098–2109, 2016, doi: 10.1109/TBME.2016.2632522.

[22] N. Ohura, R. Mitsuno, M. Sakisaka, Y. Terabe, Y. Morishige, A. Uchiyama, T. Okoshi, I. Shinji, A. Takushima, "Convolutional neural networks for wound detection: the role of artificial intelligence in wound care," *Journal of wound care*, vol. 28, no. Sup10, pp. S13–S24, 2019, doi: 10.12968/jowc.2019.28.Sup10.S13.

[23] M. Goyal, N. D. Reeves, S. Rajbhandari, N. Ahmad, C. Wang, M. H. Yap, "Recognition of ischaemia and infection in diabetic foot ulcers: Dataset and techniques," *Computers in Biology and Medicine*, vol. 117, p. 103616, 2020, doi: 10.1016/j.compbiomed.2020.103616.

[24] V. Rajinikanth, S. Kadry, P. Moreno-Ger, "Resnet18 supported inspection of tuberculosis in chest radiographs with integrated deep, lbp, and dwt features," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 2, pp. 38–46, 2023, doi: 10.9781/ijimai.2023.05.004.

[25] A. Gupta, S. Namasudra, "A novel technique for accelerating live migration in cloud computing," *Automated Software Engineering*, vol. 29, no. 1, p. 34, 2022, doi: 10.1007/s10515-022-00332-2.

[26] V. Singh, D. Jain, "A hybrid parallel classification model for the diagnosis of chronic kidney disease," vol. 8, 2023, doi: 10.9781/ijimai.2021.10.008.

[27] J.-H. Ahn, Y. Ma, S. Park, C. You, "Federated active learning (f-al): an efficient annotation strategy for federated learning," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3376746.

[28] D. Li, W. Xie, Z. Wang, Y. Lu, Y. Li, L. Fang, "Feddiff: Diffusion model driven federated learning for multi-modal and multi-clients," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, doi: 10.1109/TCSVT.2024.3407131.

[29] S. A. Rieyan, M. R. K. News, A. M. Rahman, S. A. Khan, S. T. J. Zaarif, M. G. R. Alam, M. M. Hassan, M. Ianni, G. Fortino, "An advanced data fabric architecture leveraging homomorphic encryption and federated learning," *Information Fusion*, vol. 102, p. 102004, 2024, doi: 10.1016/j.inffus.2023.102004.

[30] S. Chen, L. Li, G. Wang, M. Pang, C. Shen, "Federated learning with heterogeneous quantization bit allocation and aggregation for internet of things," *IEEE Internet of Things Journal*, vol. 11, pp. 3132–3143, 2023, doi: 10.1109/JIOT.2023.3296493.

[31] M. D. Fathima, S. J. Samuel, S. Raja, "HDDSS: An Enhanced Heart Disease Decision Support System Using RFE-ABGNB Algorithm," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 2, pp. 29-23, 2023, doi: 10.9781/ijimai.2021.10.003.

[32] H. Elayan, M. Aloqaily, M. Guizani, "Sustainability of healthcare data analysis iot-based systems using deep federated learning," *IEEE Internet of Things Journal*, vol. 9, no. 10, pp. 7338–7346, 2021, doi: 10.1109/JIOT.2021.3103635.

[33] W. Sun, S. Lei, L. Wang, Z. Liu, Y. Zhang, "Adaptive federated learning and digital twin for industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5605–5614, 2020, doi: 10.1109/TII.2020.3034674.

[34] Z. Li, X. Xu, X. Cao, W. Liu, Y. Zhang, D. Chen, H. Dai, "Integrated cnn and federated learning for covid-19 detection on chest x-ray images," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 21, pp. 835 – 845, 2022, doi: 10.1109/TCBB.2022.3184319.

[35] M. A. Rahman, M. S. Hossain, M. S. Islam, N. A. Alrajeh, G. Muhammad, "Secure and provenance enhanced internet of health things framework: A blockchain managed federated learning approach," *IEEE Access*, vol. 8, pp. 205071–205087, 2020, doi: 10.1109/ACCESS.2020.3037474.

[36] T. K. Dang, X. Lan, J. Weng, M. Feng, "Federated learning for electronic health records," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 5, pp. 1–17, 2022, doi: 10.1145/3514500.

[37] P. Baheti, M. Sikka, K. Arya, R. Rajesh, "Federated learning on distributed medical records for detection of lung nodules.," in *VISIGRAPP (4: VISAPP)*, 2020, pp. 445–451.

[38] L. Huang, A. L. Shea, H. Qian, A. Masurkar, H. Deng, D. Liu, "Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records," *Journal of biomedical informatics*, vol. 99, p. 103291, 2019, doi: 10.1016/j.jbi.2019.103291.

[39] J. Lee, J. Sun, F. Wang, S. Wang, C.-H. Jun, X. Jiang, *et al.*, "Privacy-preserving patient similarity learning in a federated environment: development and analysis," *JMIR medical informatics*, vol. 6, no. 2, p. e7744, 2018, doi: 10.2196/medinform.7744.

Sujit Kumar Das

Sujit Kumar Das received a Ph.D. from the Department of Computer Science and Engineering, National Institute of Technology Silchar, India. Currently, he works as an Assistant Professor in the Department of Computer Science and Engineering, Institute of Technical Education and Research, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, Odisha, India. His research interests include deep learning, computer vision, and medical imaging. He has published 15 SCI/SCOPUS-indexed journal articles. Dr. Das is also an active reviewer in the domain o medical imaging, deep learning, and computer vision.

Nageswara Rao Moparthi

Nageswara Rao Moparthi is a Professor at the Amrita Vishwa Vidyapeetham, Andhra Pradesh, India, in the Amrita School of Computing. Machine learning with software engineering techniques was his doctorate's major domain/specialization. His research areas include data mining, data analytics, machine learning, soft engineering and IoT. Prof. Rao has around 13 years of IT industry exposure and 7 years of teaching cum research experience. He is an active reviewer of many reputed journals and published research papers in many SCI-indexed journals. He is also an organizing committee member/TPC member of many international conferences.

Suyel Namasudra

Suyel Namasudra has received Ph.D. degree from the National Institute of Technology Silchar, Assam, India. He was a post-doctorate fellow at the International University of La Rioja (UNIR), Spain. Currently, Dr. Namasudra is working as an assistant professor in the Department of Computer Science and Engineering at the National Institute of Technology Agartala, Tripura, India. Before joining the National Institute of Technology Agartala, Dr. Namasudra was an assistant professor in the Department of Computer Science and Engineering at the National Institute of Technology Patna, Bihar, India. His research interests include blockchain technology, cloud computing, DNA computing, and information security. Dr. Namasudra has edited 8 books, 5 patents, and 87 publications in conference proceedings, book chapters, and refereed journals like IEEE TII, IEEE TNSM, IEEE TCE, IEEE T-ITS, IEEE TSC, IEEE TCSS, IEEE TCBB, ACM TOMM, ACM TOSN, ACM TALLIP, FGCS, CAEE, and many more. He is the Editor-in-Chief of the Cloud Computing and Data Science (ISSN: 2737-4092 (online)) journal. Dr. Namasudra has served as a Lead Guest Editor/Guest Editor in many reputed journals like IEEE TCE (IEEE, IF: 4.3), IEEE TBD (IEEE, IF: 7.2), ACM TOMM (ACM, IF: 3.144), MONET (Springer, IF: 3.426), CAEE (Elsevier, IF: 3.818), CAIS (Springer, IF: 4.927), CMC (Tech Science Press, IF: 3.772), Sensors (MDPI, IF: 3.576), and many more. He has also participated in many international conferences as an organizer and session chair. Dr. Namasudra is a senior member of IEEE and ACM. He has been featured in the list of the top 2% scientists in the world from 2021 to 2024. His h-index is 39.

Rubén González Crespo

Ruben Gonzalez Crespo has received PhD in Computer Science Engineering. He is the Vice-Rector of Academic and Professorate Affairs of UNIR, Spain. He is the Editor-in-Chief of the International Journal of Interactive Multimedia and Artificial Intelligence and an editorial board member of many indexed journals. His main research areas are Soft Computing, Accessibility and TEL. He is an advisory board member of the Ministry of Education, Colombia and an evaluator of the National Agency for Quality Evaluation and Accreditation of Spain (ANECA)

David Taniar

David Taniar received all his degrees (Bachelor, Master, and PhD) in Computer Science. His research expertise includes data warehousing, data management, data engineering, and data analytics. His recent book on Data Warehousing and Analytics (Springer, 2021) has been accessed more than 50 thousand times and is being used as a textbook worldwide. He has graduated more than 25 PhD students in his career. He is currently an Associate Professor at Monash University, Australia.

The Application of Deep Learning for Classification of Alzheimer's Disease Stages by Magnetic Resonance Imaging Data

Muhammad Irfan[1]*, Seyed Shahrestani[1], Mahmoud ElKhodr[2]

[1] School of Computer, Data and Mathematical Sciences, Western Sydney University, Sydney (Australia)
[2] School of Engineering and Technology, Central Queensland University, Sydney (Australia)

* Corresponding author: 19918600@student.westernsydney.edu.au

Abstract

Detecting Alzheimer's disease (AD) in its early stages is essential for effective management, and screening for Mild Cognitive Impairment (MCI) is common practice. Among many deep learning techniques applied to assess brain structural changes, Magnetic Resonance Imaging (MRI) and Convolutional Neural Networks (CNN) have grabbed research attention because of their excellent efficiency in automated feature learning of a variety of multilayer perceptron. In this study, various CNNs are trained to predict AD on three different views of MRI images, including Sagittal, Transverse, and Coronal views. This research use T1-Weighted MRI data of 3 years composed of 2182 NIFTI files. Each NIFTI file presents a single patient's Sagittal, Transverse, and Coronal views. T1-Weighted MRI images from the ADNI database are first preprocessed to achieve better representation. After MRI preprocessing, large slice numbers require a substantial computational cost during CNN training. To reduce the slice numbers for each view, this research proposes an intelligent probabilistic approach to select slice numbers such that the total computational cost per MRI is minimized. With hyperparameter tuning, batch normalization, and intelligent slice selection and cropping, an accuracy of 90.05% achieve with the Transverse, 82.4% with Sagittal, and 78.5% with Coronal view, respectively. Moreover, the views are stacked together and an accuracy of 92.21% is achived for the combined views. In addition, results are compared with other studies to show the performance of the proposed approach for AD detection.

Keywords

I. Introduction

ALZHEIMER'S Disease (AD) is the most common type of dementia. It is a irreversible, progressive, and chronic neurodegenerative disease starts with mild memory loss and possibly leading to the serious memory loss and even death [1]. AD involves parts of the brain that control thought, memory, and language. It is clinically expressed by cognitive dysfunction, amnesia, and steady loss of various brain functions and everyday living independent actions [2]. AD patients are anticipated to grow worldwide from today's figure of 47 million to 152 million by 2050 [2]. This anticipated increase will produce tremendous medical, social, and economic impacts [3], [4]. AD may cause shrinking in some areas of the human brain, reduce the brain's hippocampal size, and in some cases, lead to an enlargement in the brain ventricles [4]. Additionally, the pathogenesis of AD remains not fully explored, and the available therapies cannot reverse it or completely stop its progression. Mild Cognitive Impairment (MCI) and

Cognitive Normal (CN) tests are typically conducted by neurologists to detect AD [5]. However, these tests are challenging and complex. [6]. Studies have shown that most patients who suffer from Mild Cognitive Impairment are at risk of developing Dementia or other forms of AD. About 10–15% of people with MCI progress to AD annually [7].

Detecting AD using MCI screening is critical for successfully designing and implementing care practices and policies to counter disease deterioration. Therefore, early and stages detection is crucial to slow down the progression of the disease as it enables the development of early intervention and treatment plans [8]. Neuropathology changes in the brain help detect AD and its progression. For instance, the brain's gray matter loss has accompanied MCI and AD [6]. Typically, neurologists use clinical methodologies such as Cerebrospinal Fluid (CSF) examinations to classify AD [9]. An increase in the norepinephrine level in the CSF indicates AD progression. The CSF is usually collected directly from the brain ventricles [10]. However, CSF collection and examinations carry risks [11], [12]. Alternatively,

Cognitive Normal (CN)   Mild Cognitive Impairment (MCI)   Alzheimer's Disease (AD)
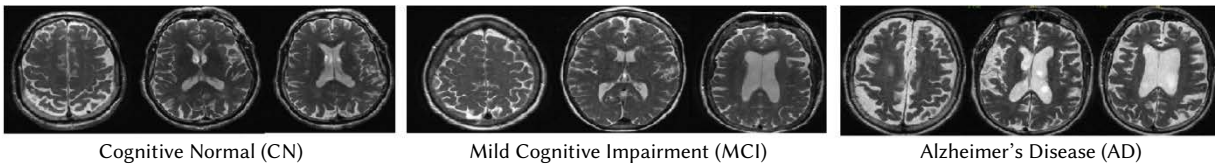
Fig. 1. Cross-sections from MRI images of CN, MCI, and AD.

non-invasive procedures such as MRI have been used by physicians to determine and assess any changes in the brain. Neurologist uses MRI scans to observe and analyze the structural changes in the brain that might be caused by AD, MCI, and CN manifestation [13]. Therefore, neuroimaging helps in visualizing the structural changes in the brain. Fig. 1 shows the changes in the brain of an AD patient. The ventricle enlargement and the changes in the hippocampal size can be observed in these MRI samples, which were taken from the ANDI database. Fig. 1 shows a comparison between images of an AD brain with cortical atrophy with the MRI images of an MCI and a CN patient. The brain texture changes with the progression of AD disease (CN to MCI to AD). Morphological changes in the texture, volume, and structure of the brain are usually used as indicators of the brain's health [14], [15].

Many studies, such as reported [16]-[26], have used neuroimaging biomarkers to predict the stages or the progress of AD. Commonly, MRI images are extensively used in all these studies due to their high resolutions and reasonable cost. Many successful machine learning frameworks have used MRI to predict AD [27], a few of them including RF (random forests) [28], SVM (support vector machine) [29], and boosting techniques [30]. Current machine learning frameworks generally involve a manual assortment of the defined ROI (regions of interest) of the patient brain based on known MRI feature representation [16]-[26], [31].

However, Manual ROI assortment can be susceptible to subjective errors [30], [32], [33]. A manual and automated ROI assortment comparison is presented in [33]. The findings demonstrate significant differences between manual and automated approaches to ROI analysis. The automated process led to a larger estimated task-related effect size. The percent of activated voxels in the automated approach was also more prominent than that of the manual approach in both lesioned and control brains and the right and left hemispheres [33].

To fill this gap, this study proposes the application of deep learning to extract signifying features from brain MRI images. The proposed method utilizes a four-layered Convolutional Neural Network (CNN) architecture to classify clinically evaluated patients with AD into people with MCI and those who are CN.

A two-dimensional (2D) CNN architecture to detect the different stages of AD is proposed in this work. The labeled data is selected from the ANDI dataset [34] and applied 2D-CNN with preprocessing to improve the detection accuracy. The MRI data to train the CNN has three different labels: AD, MCI, and CN, respectively. This work has considered T1-weighted data files. Each file contains a sagittal, coronal, and transverse view. A sample of the views is shown in Fig. 2. Noise is evident in MRI images. Therefore, preprocessing is first applied to extract the brain parts from these images. A three-layered CNN architecture with two dense layers is implemented to detect AD stages. The preprocessing pipeline in this study includes skull stripping, spatial normalization, smoothing, grey normalization, slicing, and resizing. After tuning the parameters and hyperparameters of the CNN, the prediction accuracies of AD stages are significantly improved. In addition, intelligent frame selection and batch normalization reduced the model overfitting. The main contributions of this study include:

a) Sagittal Slice   b) Coronal Slice   c) Transverse Slice
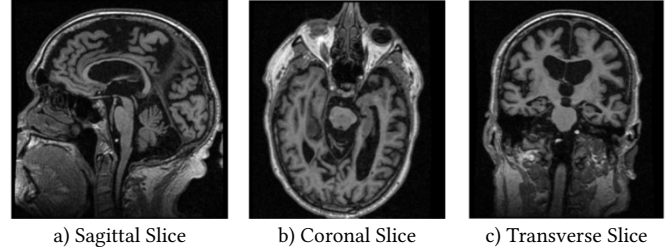
Fig. 2. Sample views of MRI images from the ADNI database.

- A 2D CNN architecture to detect the different stages of AD is proposed which is trained on the labeled data from ANDI dataset.
- Efficient preprocessing pipeline, slices selection and cropping is proposed to reduce the input data size and avoid model's underfitting.
- The proposed model is trained on three views of MRI, separately and combined.
- A wide set of comparison is performed between different views and the recent state-of-the-art literature.

The remaining of this paper is organized as follows. Section II presents the background of this work and its related studies. Our proposed methodology for classifying AD stages is explained in Section III. Section IV discusses the obtained experimental results and outcomes. Finally, the last section gives the conclusions of this work.

II. Related Studies

Deep learning approaches attempt to imitate the human brain by utilising CNNs, RNNs, stacked auto-encoder, and deep belief networks (DBNs) [35], [36], [37], [38]. They transform low-level features available in the data to build an abstract high-level representation of the learning systems [39]. A dual-tree complex wavelet transform-based method in [40] extracts features from the input, followed by classification with FDNN (feedforward neural network). CNN is a deep multilayer artificial neural network (ANN) composed of convolutional layers, allowing a model to extract feature maps learned from the product of inputs and kernels, thereby detecting the patterns. Moreover, CNNs have shown high accuracy in feature classification [41]-[43]. In the segmentation applications, CNNs outperformed other methodologies such as SVM and logistic regression, which showed less intrinsic feature extraction capabilities [44]. CAD (Computer-Aided Diagnosis) systems built on CNNs successfully detect neurodegenerative diseases [45]. CNN architectures, including the ResNet and GoogleNet, have been successfully used in differentiating the healthy from AD and MCI [46]. LeNet-5 CNN architecture differentiates AD from the NC brain [47]. A deep supervised adaptive 3D-CNN in [48] predicted AD by stacking 3D Convolutional autoencoders without stripping the skull structure. ResNet-152 in [41] obtained highly-discriminative features to detect the stage of the disease progression (AD, MCI, and CN) using neuroimaging data taken from the ADNI database. The study in [49] has used transfer learning and VGG-16 pre-trained architecture for multiclass AD classification on AD, MCI, and CN. The study in [50] implemented 3D-ResNet-18 with data augmented Resnet-18 for feature extraction to classify AD stages accurately. ResNet-18 architecture

was modified in [51] for the binary AD classification: CN vs. AD, CN vs. MCI, AD vs. MCI, and CN vs. MCI. The Transfer Learning scheme is used in [52] for the three-way classification (AD, CN, MCI) of MRI images, implemented in three pre-trained CNNs, including ResNet-18, ResNet-50, and ResNet-101, respectively. A 2D-CNN architecture in [53] used ResNet-50 with diverse activations and batch normalization to classify brain slices into NC, MCI, and AD.

SegNet can classify patients' AD stages using extracted morphological local features from the brain [54]. Resnet-101 also attempted to classify AD, MCI, and CN stages. In [55], A 3D-CNN used a classifier to differentiate the CN and AD using brain MRI images. Using the ADNI dataset, a probability-based CNNs fusion in [47] used DenseNet to detect AD stages. A 3-D Net-121 with a 70% dropout rate is shown to detect the AD stages [56]. A layer-wise Transfer Learning using VGG-19 in [57] discriminated the CN, early MCI, late MCI, and AD. Another Transfer Learning method was presented in [58], which recommended VGG-16 to accurately classify brain MRI slices into CN, MCI, and AD. A pre-trained AlexNet in [59] extracted significant features from the MRI images to classify the AD. Another fine-tuned pre-trained AlexNet, presented in [60] used Transfer Learning to classify the MRI images. In addition, a modified AlexNet in [61] with the parameters adjustment discriminated AD stages. In [62], various pre-trained architectures were utilized after fine-tuning the Transfer Learning approach for CN, MCI, and AD classification from the ADNI dataset. For the AD and MCI prediction, an ensemble of densely-connected 3D-CNNs is suggested in [63] for improving usage of extracted features. The CNN topologies for the binary classification (AD/MCI or MCI/CN) are proposed in [64] by integrating freezing characteristics engaged from ImageNet dataset. Usullay MRI images are used with one or few views without removing the redundant information and noise. In addition, no preprocessing is generally applied for neural networks to learn better. After styding the related research, the issues addressed by preprocessing the MRI images along with noise removal for efficient feature learning. Many denoisiyng methods can be applied such as in [65]-[67].

A 2D CNN architecture to detect different stages of AD is proposed in this work. The CNN architecture is fine-tuned on the dataset, which is preprocessed by a feature engineering method. After parameters and hyperparameters tuning the CNN, AD stages' prediction accuracies are significantly improved. The intelligent frame selection and batch normalization reduced model overfitting. The MRI data to train the CNN has three different labels: AD, MCI, and CN, respectively. This work has considered T1-weighted data files. Each file contains a sagittal, coronal, and transverse view.

III. Methodology

A. Dataset

This study uses the MRI data from the Alzheimer's disease Neuroimaging Initiative (ADNI) database. ADNI is a labeled dataset that has three different labels, i.e., Alzheimer's disease (AD), Mild Cognitive Impairment (MCI), and Cognitive Normal (CN). This dataset contains data about the multiple visits of the same patient during the trial. There are 2182 NIFTI files (3D view MRI images). Each NIFTI file contains a single subject's sagittal, coronal, and transverse views. There is almost 200+ sequenced frames of all three views of the MRI images. Sample views of some of the MRI images taken from the ADNI database are presented in Fig. 2. Since the initial views are noisy and difficult to process, therefore, brain part has been extracted from the MRI to allow further processing. These details of this process are provided in Section B. The dataset distribution gender and label, along with the statistics, age, and the number of visits, are given in Table I.

TABLE I. The Dataset Distribution

Male Patients	1279
Female Patients	930
Cognitive Normal (CN)	748
Mild Cognitive Impairment (MCI)	981
Alzheimer's Disease (AD)	453
Average Age of Patients	76.23
Average Patients Visits	4.10
Age Standard Deviation	6.80

B. MRI Preprocessing

The T1-Weighted MRI data from the ANDI in NIFTI format was preprocessed using the CAT12 toolkit of SPM12 toolbox (MATLAB third party toolbox) with default settings. The preprocessing pipeline includes skull stripping, spatial normalization and smoothing such that after preprocessing, all MRI images follow the dimension $(121 \times 145 \times 121)$, that is $(X \times Y \times Z)$ with a spatial resolution of $(1.5 \times 1.5 \times 1.5)$ mm3/voxel. In addition, all MRI images, including each voxel value, were normalized in terms of signal intensity. The original value was divided by the actual maximal value of the MRI image. This normalization yields values in the range of 0 and 1. The resultant views after preprocessing the pipeline are shown in Fig. 3. The 3D-MRI $(121 \times 145 \times 121)$, which is the number of sagittal, coronal, and transverse views, were acquired via re-slicing, i.e., (145×121), (121×121), and (121×145), respectively. All the 2D slices were resized to (145×145) after edge padding and zero filling. After resizing, each 2D slice was squared, whereas the central and spatial resolution of the reformatted MRI image remained unchanged.
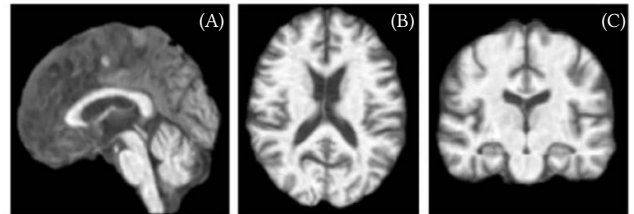
Fig. 3. MRI Preprocessed Views; (A) Sagittal Slice, (B) Coronal Slice, and (C) Transverse Slice.

1. Skull Stripping

A skull stripping method is integral in brain image processing applications [68]. It acts as a preliminary step in numerous medical ML applications as it increases the speed and accuracy of diagnosis manifold [69]. It removes non-cerebral tissues like the skull, scalp, and dura from brain images. Adaptive Probability Region-Growing (APRG) is a method that refines the probability maps by region-growing techniques [64]. This is currently the method with the most accurate and reliable results. This research has removed the skull from MRI data using the APRG method.

2. Spatial Normalization

Human brains differ in size and shape, and one goal of spatial normalization is to deform the human brain scans, so one location in one subject's brain scan corresponds to the same location in another subject's brain scan. More specifically, images from different subjects must be transformed spatially so that they all reside in the same coordinate system, with anatomically corresponding regions being in similar locations. Spatial normalization is a particular form of image registration that maps a subject's MRI image to a reference brain space to allow comparisons across subjects with varied brain morphologies [70]. This research used Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra (DARTEL) registrations and its existing templates for spatial registration. Furthermore, an optimized

shooting approach is applied that uses an adaptive threshold and lower initial resolutions to obtain a good tradeoff between accuracy and calculation time by selecting the first of the six images (iterations) of a DARTEL template similar to work conducted in [65].

3. Smoothing

Smoothing is used to remove the different noises from the MRI frames. Then, the Gaussian filter is applied to the MRI data to reduce the noise. The images are shown in Fig. 3, which depicts the finalized preprocessed frames of all three views.

C. Proposed Features Engineering

After preprocessing, there are almost 150 slices per view per MRI image. These large numbers of frames require substantial computation power to train a CNN model on them. Moreover, data redundancy would cause the CNN to be overfitting. Therefore, reducing the number of frames of each view is essential to reduce the total computational power needed to process each MRI. To address this challenge, a recent research has randomly selected 40 sagittal slices, 50 coronal slices, and 33 transverse slices, i.e., 123 slices of a subject's 3D brain image [66]. However, the random selection of frames is not convincing as it is unknown which frame contains more information. Random selection can lead to loss of information. To fill this gap, this research relied on a new method that used statistical analysis when selecting the important frames. Firstly, several informative pixels are calculated. If a slice has less than a threshold value of informative pixels, these slices were discarded, and the remaining frames are selected. The formula used for calculating the number of informative pixels is given in (1):

$$I = 1 - \frac{N_0}{H_I \times W_I} \qquad (1)$$

Where N_0 is the number of zeros in an image, H_I is the Height of the image, and W_I is the width of the image. This study has selected the highest 40 informative frames of each view, resulting in 120 frames per patient. The process of statistical selection of frames has resulted in reducing the computational complexity of MRI features selection process. Given that every single frame contained various sizes of the informative region. A generic average windows size was calculated for all the patients and all the three views (sagittal, coronal, and transverse) using the proposed algorithm. The subsequent slices have further reduced the computational complexity of the MRI file processing. A sample transverse view after slice cropping is shown in Fig. 4.

Fig. 4. Sample Transverse View of Slices.
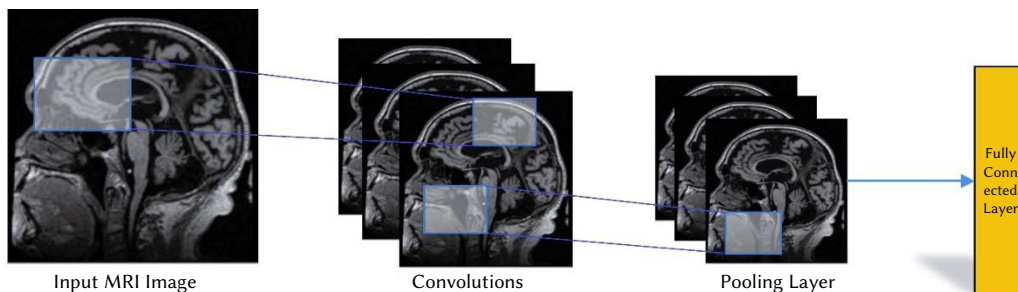
Algorithm for Generic Windows Size
for slice in ISS (Intelligently Selected Slices):
    for patient in patients:
        find top, bottom, left, and right first zero vector
        drop the zero valued vectors from frame
average the cropped image (*windows size*)

D. Convolutional Neural Network (CNN)

The CNN neural model has recently gained significant research attention with remarkable success in recognizing images [42]. The input images move through a chain of convolution layers with CNN, including filtering, pooling, and fully connected dense layers. In addition, the softmax activation function is usually applied for a probabilistic classification of the images between 0 and 1, making CNN appropriate for image feature representation learning. In CNN, a convolution layer contains the extraction and mapping of features. During the feature extraction, all neurons are connected to the local accessible fields of the higher layer for local feature extraction. After the local feature extraction, a spatial relationship with other features is concluded.

On the other hand, convolution operations are applied to the input data using learnable filters (kernels) to produce a feature map during the feature mapping. Multiple feature maps can be computed with a chain of filters. In this manner, the CNN parameters are tuned and can be effectively reduced. After the convolutional layer, the max-pooling layer executes a down-sampling operation in addition to the spatial dimensions. Such a distinctive dual-feature extraction scheme can successfully moderate the feature resolution. The activations usually use nonlinear functions such as the sigmoid, tanh, ReLU, and Leaky ReLU. To accelerate the learning and prevent overfitting of the proposed model, pooling layers were integrated into the CNN. This layer reduces the samples extracted from the data, thereby reducing the spatial information. Average pooling and max-pooling are the prominent pooling schemes. The FC (fully-connected) layer is similar to the Artificial Neural Network (ANN). Its task is to set a path for effective detection. An example demonstration of the CNN model's convolutional, pooling, and fully connected layers are shown in Fig. 5.

This study mainly used CNN with the following architecture to recognize 2D MRI images. The preprocessed MRI image was fed into the CNN model as feature extraction and mapping vectors. Then, the max-pooling layer learns the features from the training data. This process improves the effectiveness of CNN instead of manually extracting the features. The CNN was trained by applying the learnable filters and convolutional operations. Using a local weight distribution has significantly reduced the complexity of the model. The format for CNN with 3D input data follows (Width × Height × No. of Frames). All three views of a single MRI image were treated separately. For each view, an individual CNN was trained. The model architecture for 3D inputs is provided in Table II. The model configurations are provided in Table III.
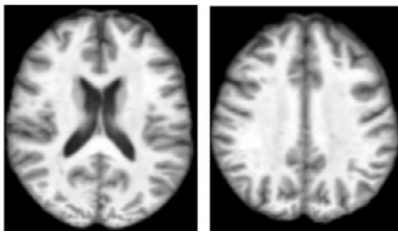

Fig. 5. Example demonstration of the convolutional, pooling, and fully connected layers of the CNN model.

TABLE II. Layers and Parameters in CNN Model for MRI Images

Layers	Output Shape	Para#	Layers	Output Shape	Para#	Layers	Output Shape	Para#
			Conv2D	(13, 10, 64)	18496	Conv2D	(4 3, 128)	73856
Conv2D	(41, 32, 32)	11552	BN	(13, 10, 64)	256	BN	(4 3, 128)	512
L-ReLU	(41, 32, 32)	0	L-ReLU	(13, 10, 64)	0	L-ReLU	(4 3, 128)	0
			Maxpool	(13, 10, 64)	0	Maxpool	(4 3, 128)	0
Flatten Layer: (None, 1536), Para# = 0								
Dense Layer: (None, 100), Para# = 153700								
Dense Layer: (None, 3), Para# = 303								
Total Trainable Para#: 258,291								

(BN= Batch Normalization, Conv2D= 2D Convolution, L-ReLU= Leaky ReLU, Maxpool= Maxpooling layer)

TABLE III. Model Configurations Table

Parameter	Configuration	
Learning Rate	Initial Value	0.001
	Nature	Timely Decreasing (Adaptive)
	Reduction Factor	0.1
	Minimum Possible Value	0.00001
	Reduction Monitoring	Validation Accuracy
	Patient to Reduction	2 times
Stopping Criteria	Stopping Monitoring	Validation Accuracy
	Patience to Stop	20 times
	Initial Learning Rate	0.001
Weights	Trainable	Yes
	Initial Weights	Random
Training	Optimizer	Adam
	Loss	Categorical Cross-Entropy
	Maximum Possible Epochs	Infinite
	Batch Size	32
	Validation Split	15%
	Performance Metric	Accuracy
		Loss
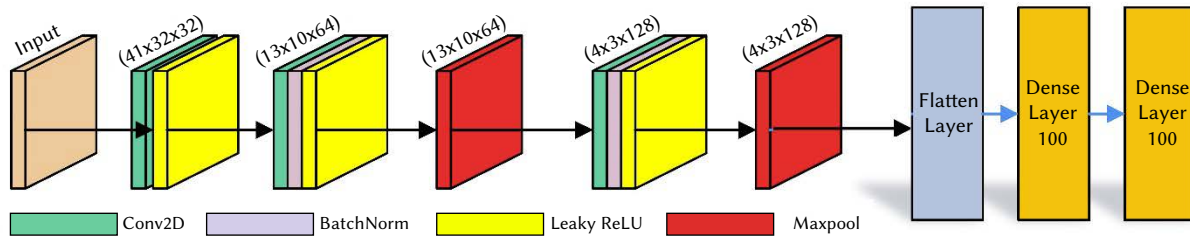
Fig. 6. The architecture of the layers used in the CNN.

The CNN model proposed in this study contains three convolutional layers with different filter numbers, sizes, and strides, two max-pooling layers, and two fully-connected dense layers. Leaky activation and batch normalization are used after the convolutional layers. The architecture of the layers used in the CNN is shown in Fig. 6. The first convolutional layer contains 32 (3×3) size filters and stride 1. Similarly, the second and third convolutional layer contains 64 and 128 filters of size (3×3) and stride 1. All pooling layers use a max-pooling scheme with the pooling window size of (2×2) and strides 2. The last max-pooling layer's output is passed through the flattening layer and converted 2D data into 1D. The output of the flattened layer is fed to the fully-connected dense layer with 100 neurons using softmax as the activation function. The fully-connected layer is an Artificial Neural Network (ANN) based- architecture. ADAptive Moment (ADAM) estimation is used to optimize the learning weights of model, which is extended version of stochastic gradient decent [68]. The learning rate

and momentum are fixed to 0.0001, 0.9, and the binary cross-entropy loss function in the CNN model training. If validation accuracy is not improving and patience to stop reaches to maximum point, the model stops at that point, as provided in Table III.

IV. Experimental Results and Discussions

The intelligent frames selection dropped the initial frames with more than 50% zeros. After intelligent frame selection, the frames were normalized between 0 and 1. Finally, the preprocessed data is trained and tested on the same CNN model (as explained in section 3.3). The methodology includes three different strategies. Firstly, different MRI views were trained on the same CNN. This research first presents transverse slices' results where the input frame size was (123×98). Initially, an accuracy of 80.21% is achieved with the transverse view. However, after tuning the parameters and hyperparameters of CNN

model, the prediction accuracy improved from 80.21% to 90.5±1.5%. In addition, an intelligent frame selection and batch normalization reduced the model overfitting. Also, a precision of 87.2±1.2%, Recall with 90.7±1%, and F1 with 90.7±1% F1 measures are achieved with transverse views. Secondly, the results of the sagittal slices where the input frame size to CNN was (106×123) were noted. With the proposed probability-based frame selection and features engineering, the achieved accuracy was 80.5±2.5% for sagittal views.

Moreover, other measures achieved are precision of 80.5±2.5%, Recall with 82.3±2.7%, and F1 score with 81.5±3.1%. Lastly, the results of the coronal slices, where the input frame size was (104×97), were compiled with intelligent frame selection and feature engineering. It achieved an accuracy of 78.5±4.5% for the coronal views. Table IV shows the performance measures in terms of accuracy, precision, recall, and F1 scores for three MRI views, sagittal, coronal, and transverse, respectively. The results indicate that better performance was achieved with transverse views, and the accuracy has improved from 78.5 ± 4.5% to 90.5 ± 1.5%, whereas precision improved from 80.5 ± 2.5% to 87.2 ± 1.2%. This research work has combined all views and trained the same CNN architecture. With this strategy, an accuracy of 92.21% has been achieved. Table V shows all the corresponding measures for the three combined views. The CNN model has been trained and validated for 50 epochs and achieved 90.41% testing accuracy on transverse view.

TABLE IV. CNN's Performance Analysis for Three Separate Views

View	Transverse	Sagittal	Coronal
Accuracy	90.5 ± 1.5	82.4 ± 2.9	78.5 ± 4.5
Precision	87.2 ± 1.2	80.5 ± 2.5	77.3 ± 3.3
Recall	90.7 ± 1	82.3 ± 2.7	79.1 ± 2.8
F1-score	90 ± 1.3	81.5 ± 3.1	77.8 ± 3.1

TABLE V. Performance Analysis of CNN Trained on All Views

Combined Views			
Accuracy	Precision	Recall	F1-Score
92.21 ± 1	89.47 ± 1.3	91.05 ± 2.7	90.25 ± 3.1

Additionally, to further evaluate the performance of the proposed CNN model, the model with frame selection was compared with the PCA+SVM [71], 3D-SENet and CNN+EL 3D-SENet [66]. As a central element of CNN, the convolution operation enables networks to obtain informative feature representation by combining spatial and channel-wise information within local fields. The results in terms of the accuracy of the different models are given in Table VI. It can be observed that CNN with the proposed probabilistic frame selection for early AD detection is more accurate and robust than the PCA+SVM, 3D-SENet, and CNN+EL. The model's accuracy improved from (71.33 ± 0.29)% with PCA+SVM to (83.33 ± 2.96) % with the proposed approach. Also, the accuracy is increased from (75.11 ± 0.23)% and (75.11 ± 0.60) % with 3D-SENet and CNN+EL to (83.33 ± 2.96)% with the proposed model.

TABLE VI. Performance of the Proposed CNN With Other Models

Models	Accuracy of Models
PCA+SVM	71.33 ± 0.29
CNN+EL	75.11 ± 0.60
3D-SENet	75.11 ± 0.23
Proposed	83.33 ± 2.96

The transfer learning with pre-trained models, such as the AlexNet, VGG, and GoogleNet, was also examined against the proposed approach. Most of the classification-prediction problems in medical imaging have been implemented with SVM. Table VII compares the

SVM and other models to investigate the performance. The trained model was the most convenient tool for the AD detection prediction problem based on classification accuracy and prediction responses. The SVM classifier presented the result with (70~80) % accuracy. However, CNN models provide a prediction accuracy of (80~90) %, as represented in Fig. 7.

TABLE VII. Performance Analysis of the Proposed CNN With State-of-the-Art Models

Models	Accuracy of Models
Deep learning using AlexNet	85.14%
Deep learning using VGG16	88.92%
Deep learning using VGG19	90.02%
Deep learning using GoogLeNet	87.29%
Support Vector Machine (SVM)	74.25%
Proposed	92.21 ± 2.96
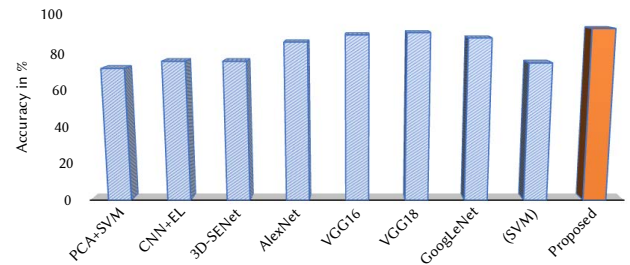
Fig. 7. Overall Accuracy of SVM and CNN-basedModels.

V. Conclusions

This research aimed to deploy unconventional deep learning methods to determine whether they can extract helpful Alzheimer's disease biomarkers from Magnetic Resonance Imaging and classify brain images into Alzheimer's disease, Mild Cognitive Impairment, and Normal Cognitive groups. The T1-Weighted MRI data from the ANDI database in NIFTI format was preprocessed, but many MRI slices required a massive computational cost. Therefore, it was essential to reduce the slice numbers for three MRI views to minimize the total computational cost. Probabilistic frame selection was proposed to address the problem, which has improved the overall CNN accuracy from 80.21% to 90.5±1.5. This study first trained CNNs on three different MRI views (transverse, sagittal, and coronal). During this set of experiments, the proposed approach achieved a better accuracy of (90.5 ± 1.5) and a precision of (87.2 ± 1.2) for transverse views.

Furthermore, all three views were combined and tested with CNN using MRI scans. This proposed approach improved the performance and achieved (92.21 ± 1) accuracy and (89.47 ± 1.3) precision. With intelligent frame selection, the trained CNN model achieved an accuracy of 92.21%, which is significantly better than similar models. It is observed that the absolute accuracy has increased from 71.33% (PCA+SVM), 75.11% (CNN+EL), and 75.11% (3D-SDNet) to 83.33% with the proposed CNN approach. The transfer learning with pre-trained models, such as the AlexNet, VGG, and GoogleNet, was also examined. The CNN models obtained the results in terms of accuracy of (80~90) %, which is higher than SVM classifier, which produced results with an accuracy of (70~80) %. The end-to-end implementation of CNN to classify AD, MCI, and CN groups in three different MRI views reflects the identification of distinctive elements in brain images. In this context, the proposed approach represents a promising tool in finding the biomarkers helping the early AD detection, eventually taking critical and successful care policies to counter deterioration in the disease.

This study includes on T1-weighted MRI scans to detect the different stages of AD. Although with such arrangement, the model achieved the state-of-the art results surpassing the related studies, but furher improvements can be achieved if T2-weighted scans are also utilized. In addition, microarray gene expression data can be used to classify the disease.

References

[1] M. G. Ulep, S. K. Saraon, and S. McLea, "Alzheimer disease," The Journal for Nurse Practitioners, vol. 14, pp. 129-135, 2018.

[2] C. Patterson, "World Alzheimer report 2018," 2018.

[3] J. Wiley, "Alzheimer's disease facts and figures," Alzheimers Dement, vol. 17, pp. 327-406, 2021.

[4] A. Fahimi, M. Noroozi, & A. Salehi, (2021). Enlargement of early endosomes and traffic jam in basal forebrain cholinergic neurons in Alzheimer's disease. Handbook of Clinical Neurology, 179, 207-218.

[5] J. C. Morris and J. L. Price, "Pathologic correlates of nondemented aging, mild cognitive impairment, and early-stage Alzheimer's disease," Journal of Molecular Neuroscience, vol. 17, pp. 101-118, 2001.

[6] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. Feng, "Early diagnosis of Alzheimer's disease with deep learning," in 2014 IEEE 11th international symposium on biomedical imaging (ISBI), 2014, pp. 1015-1018.

[7] S. Afzal, M. Maqsood, U. Khan, I. Mehmood, H. Nawaz, F. Aadil & Y. Nam, "Alzheimer Disease Detection Techniques and Methods: A Review," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, pp. 26-39, 2021.

[8] G. Karas, P. Scheltens, S. A. Rombouts, P. J. Visser, R. A. van Schijndel, N. C. Fox, et al., "Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease," Neuroimage, vol. 23, pp. 708-716, 2004.

[9] K. Blennow, "Cerebrospinal fluid protein biomarkers for Alzheimer's disease," NeuroRx, vol. 1, pp. 213-225, 2004.

[10] R. Elrod, E. R. Peskind, L. DiGiacomo, K. I. Brodkin, R. C. Veith, and M. A. Raskind, "Effects of Alzheimer's disease severity on cerebrospinal fluid norepinephrine concentration," The American journal of psychiatry, vol. 154, pp. 25-30, 1997.

[11] O. Hansson, S. Lehmann, M. Otto, H. Zetterberg, and P. Lewczuk, "Advantages and disadvantages of the use of the CSF Amyloid β (Aβ) 42/40 ratio in the diagnosis of Alzheimer's Disease," Alzheimer's research & therapy, vol. 11, pp. 1-15, 2019.

[12] S. Das and S. Basu, "Multi-targeting strategies for Alzheimer's disease therapeutics: pros and cons," Current Topics in Medicinal Chemistry, vol. 17, pp. 3017-3061, 2017.

[13] S. J. Teipel, M. Grothe, S. Lista, N. Toschi, F. G. Garaci, and H. Hampel, "Relevance of magnetic resonance imaging for early detection and diagnosis of Alzheimer disease," Medical Clinics, vol. 97, pp. 399-424, 2013.

[14] G. H. Weissberger, J. V. Strong, K. B. Stefanidis, M. J. Summers, M. W. Bondi, and N. H. Stricker, "Diagnostic accuracy of memory measures in Alzheimer's dementia and mild cognitive impairment: a systematic review and meta-analysis," Neuropsychology review, vol. 27, pp. 354-388, 2017.

[15] S. Lorio, F. Kherif, A. Ruef, L. Melie-Garcia, R. Frackowiak, J. Ashburner, et al., "Neurobiological origin of spurious brain morphological changes: A quantitative MRI study," Human brain mapping, vol. 37, pp. 1801-1815, 2016.

[16] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, et al., "Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database," neuroimage, vol. 56, pp. 766-781, 2011.

[17] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, and A. s. D. N. Initiative, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," Neuroimage, vol. 55, pp. 856-867, 2011.

[18] T. Tong, Q. Gao, R. Guerrero, C. Ledig, L. Chen, D. Rueckert, et al., "A novel grading biomarker for the prediction of conversion from mild cognitive impairment to Alzheimer's disease," IEEE Transactions on Biomedical Engineering, vol. 64, pp. 155-165, 2016.

[19] T. Tong, R. Wolz, Q. Gao, R. Guerrero, J. V. Hajnal, D. Rueckert, et al., "Multiple instance learning for classification of dementia in brain MRI," Medical image analysis, vol. 18, pp. 808-818, 2014.

[20] R. Guerrero, R. Wolz, A. Rao, D. Rueckert, and A. s. D. N. Initiative, "Manifold population modeling as a neuro-imaging biomarker: application to ADNI and ADNI-GO," NeuroImage, vol. 94, pp. 275-286, 2014.

[21] H.-I. Suk, S.-W. Lee, D. Shen, and A. s. D. N. Initiative, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," NeuroImage, vol. 101, pp. 569-582, 2014.

[22] B. Cheng, M. Liu, D. Zhang, B. C. Munsell, and D. Shen, "Domain transfer learning for MCI conversion prediction," IEEE Transactions on Biomedical Engineering, vol. 62, pp. 1805-1817, 2015.

[23] S. F. Eskildsen, P. Coupé, V. S. Fonov, J. C. Pruessner, D. L. Collins, and A. s. D. N. Initiative, "Structural imaging biomarkers of Alzheimer's disease: predicting disease progression," Neurobiology of aging, vol. 36, pp. S23-S31, 2015.

[24] F. Li, L. Tran, K.-H. Thung, S. Ji, D. Shen, and J. Li, "A robust deep model for improved classification of AD/MCI patients," IEEE journal of biomedical and health informatics, vol. 19, pp. 1610-1616, 2015.

[25] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, et al., "Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease," IEEE Transactions on Biomedical Engineering, vol. 62, pp. 1132-1140, 2014.

[26] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, J. Tohka, and A. s. D. N. Initiative, "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects," Neuroimage, vol. 104, pp. 398-412, 2015.

[27] J. M. Mateos-Pérez, M. Dadar, M. Lacalle-Aurioles, Y. Iturria-Medina, Y. Zeighami, and A. C. Evans, "Structural neuroimaging as clinical predictor: A review of machine learning applications," NeuroImage: Clinical, vol. 20, pp. 506-522, 2018.

[28] E. E. Tripoliti, D. I. Fotiadis, and M. Argyropoulou, "A supervised method to assist the diagnosis and monitor progression of Alzheimer's disease using data from an fMRI experiment," Artificial intelligence in medicine, vol. 53, pp. 35-45, 2011.

[29] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based tissue classification of MR images of the brain," IEEE transactions on medical imaging, vol. 18, pp. 897-908, 1999.

[30] C. Hinrichs, V. Singh, L. Mukherjee, G. Xu, M. K. Chung, S. C. Johnson, et al., "Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset," Neuroimage, vol. 48, pp. 138-149, 2009.

[31] N. Arunkumar, M. A. Mohammed, S. A. Mostafa, D. A. Ibrahim, J. J. Rodrigues, and V. H. C. de Albuquerque, "Fully automatic model-based segmentation and classification approach for MRI brain tumor using artificial neural networks," Concurrency and Computation: Practice and Experience, vol. 32, p. e4962, 2020.

[32] F. Li, M. Liu, and A. s. D. N. Initiative, "Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks," Computerized Medical Imaging and Graphics, vol. 70, pp. 101-110, 2018.

[33] K. A. Garrison, C. Rogalsky, T. Sheng, B. Liu, H. Damasio, C. J. Winstein, et al., "Functional MRI preprocessing in lesioned brains: manual versus automated region of interest analysis," Frontiers in Neurology, vol. 6, p. 196, 2015.

[34] N. R. de Vent, J. A. Agelink van Rentergem, B. A. Schmand, J. M. Murre, A. Consortium, and H. M. Huizenga, "Advanced Neuropsychological Diagnostics Infrastructure (ANDI): A normative database created from control datasets," Frontiers in Psychology, vol. 7, p. 1601, 2016.

[35] Y. LeCun, "LeNet-5, convolutional neural networks," URL: http://yann. lecun. com/exdb/lenet, vol. 20, p. 14, 2015.

[36] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," Journal of machine learning research, vol. 11, 2010.

[37] G. E. Hinton, "Deep belief networks," Scholarpedia, vol. 4, p. 5947, 2009.

[38] Z.-J. Sun, L. Xue, Y.-M. Xu, and Z. Wang, "Overview of deep learning," Jisuanji Yingyong Yanjiu, vol. 29, pp. 2806-2810, 2012.

[39] M. Sharma, P. Sharma, R. B. Pachori, and U. R. Acharya, "Dual-tree complex wavelet transform-based features for automated alcoholism identification," International Journal of Fuzzy Systems, vol. 20, pp. 1297-1308, 2018.

[40] A. Farooq, S. Anwar, M. Awais, and S. Rehman, "A deep CNN based multiclass classification of Alzheimer's disease using MRI," in 2017 IEEE International Conference on Imaging systems and techniques (IST), 2017, pp. 1-6.

[41] C. Haarburger, M. Baumgartner, D. Truhn, M. Broeckmann, H. Schneider, S. Schrading, et al., "Multi scale curriculum CNN for context-aware breast MRI malignancy classification," in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019, pp. 495-503.

[42] B. Khagi and G.-R. Kwon, "3D CNN design for the classification of Alzheimer's disease using brain MRI and PET," IEEE Access, vol. 8, pp. 217830-217847, 2020.

[43] S. Deepak and P. Ameer, "Automated categorization of brain tumor from mri using cnn features and svm," Journal of Ambient Intelligence and Humanized Computing, vol. 12, pp. 8357-8369, 2021.

[44] A. Shakarami, H. Tarrah, and A. Mahdavi-Hormat, "A CAD system for diagnosing Alzheimer's disease using 2D slices and an improved AlexNet-SVM method," Optik, vol. 212, p. 164237, 2020.

[45] A. Ashraf, S. Naz, S. H. Shirazi, I. Razzak, and M. Parsad, "Deep transfer learning for alzheimer neurological disorder detection," Multimedia Tools and Applications, vol. 80, pp. 30117-30142, 2021.

[46] L. M. Heising and S. Angelopoulos, "Operationalizing fairness in medical AI adoption: Detection of early Alzheimer's Disease with 2D CNN," BMJ Health & Care Informatics, 2022.

[47] Y. Huang, J. Xu, Y. Zhou, T. Tong, X. Zhuang, and A. s. D. N. Initiative, "Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network," Frontiers in neuroscience, p. 509, 2019.

[48] D. Ebrahim, A. M. Ali-Eldin, H. E. Moustafa, and H. Arafat, "Alzheimer Disease Early Detection Using Convolutional Neural Networks," in 2020 15th International Conference on Computer Engineering and Systems (ICCES), 2020, pp. 1-6.

[49] A. Ebrahimi, S. Luo, R. Chiong, and A. s. D. N. Initiative, "Deep sequence modelling for Alzheimer's disease detection using MRI," Computers in Biology and Medicine, vol. 134, p. 104537, 2021.

[50] M. Odusami, R. Maskeliūnas, R. Damaševičius, and T. Krilavičius, "Analysis of Features of Alzheimer's Disease: Detection of Early Stage from Functional Brain Changes in Magnetic Resonance Images Using a Finetuned ResNet18 Network," Diagnostics, vol. 11, p. 1071, 2021.

[51] D. Prakash, N. Madusanka, S. Bhattacharjee, C.-H. Kim, H.-G. Park, and H.-K. Choi, "Diagnosing Alzheimer's disease based on multiclass MRI scans using transfer learning techniques," Current medical imaging, vol. 17, pp. 1460-1472, 2021.

[52] K. S. Yadav and K. P. Miyapuram, "A Novel Approach Towards Early Detection of Alzheimer's Disease Using Deep Learning on Magnetic Resonance Images," in International Conference on Brain Informatics, 2021, pp. 486-495.

[53] P. Buvaneswari and R. Gayathri, "Deep learning-based segmentation in classification of Alzheimer's disease," Arabian Journal for Science and Engineering, vol. 46, pp. 5373-5383, 2021.

[54] H. Parmar, B. Nutter, R. Long, S. Antani, and S. Mitra, "Spatiotemporal feature extraction and classification of Alzheimer's disease using deep learning 3D-CNN for fMRI data," Journal of Medical Imaging, vol. 7, p. 056001, 2020.

[55] B. Solano-Rojas, R. Villalón-Fonseca, and G. Marín-Raventós, "Alzheimer's disease early detection using a low cost three-dimensional densenet-121 architecture," in International Conference on Smart Homes and Health Telematics, 2020, pp. 3-15.

[56] A. Mehmood, S. Yang, Z. Feng, M. Wang, A. S. Ahmad, R. Khan, et al., "A transfer learning approach for early diagnosis of Alzheimer's disease on MRI images," Neuroscience, vol. 460, pp. 43-52, 2021.

[57] R. Jain, N. Jain, A. Aggarwal, and D. J. Hemanth, "Convolutional neural network based Alzheimer's disease classification from magnetic resonance brain images," Cognitive Systems Research, vol. 57, pp. 147-159, 2019.

[58] F.-P. An, "Medical image classification algorithm based on weight initialization-sliding window fusion convolutional neural network," complexity, vol. 2019, 2019.

[59] M. Maqsood, F. Nazir, U. Khan, F. Aadil, H. Jamal, I. Mehmood, et al., "Transfer learning assisted classification and detection of Alzheimer's disease stages using 3D MRI scans," Sensors, vol. 19, p. 2645, 2019.

[60] H. Acharya, R. Mehta, and D. K. Singh, "Alzheimer Disease Classification Using Transfer Learning," in 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1503-1508.

[61] R. A. Hazarika, D. Kandar, and A. K. Maji, "An experimental analysis of different deep learning based models for Alzheimer's disease classification using brain magnetic resonance images," Journal of King Saud University-Computer and Information Sciences, 2021.

[62] H. Wang, Y. Shen, S. Wang, T. Xiao, L. Deng, X. Wang, et al., "Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease," Neurocomputing, vol. 333, pp. 145-156, 2019.

[63] S. Naz, A. Ashraf, and A. Zaib, "Transfer learning using freeze features for Alzheimer neurological disorder detection using ADNI dataset," Multimedia Systems, pp. 1-10, 2021.

[64] R. Pohle and K. D. Toennies, "Segmentation of medical images using adaptive region growing," in Medical Imaging 2001: Image Processing, 2001, pp. 1337-1346.

[65] J. Ashburner, "A fast diffeomorphic image registration algorithm," Neuroimage, vol. 38, pp. 95-113, 2007.

[66] D. Pan, A. Zeng, L. Jia, Y. Huang, T. Frizzell, and X. Song, "Early detection of Alzheimer's disease using magnetic resonance imaging: a novel approach combining convolutional neural networks and ensemble learning," Frontiers in neuroscience, vol. 14, p. 259, 2020.

[67] A. Laishram, & K. Thongam, "Automatic Classification of Oral Pathologies Using Orthopantomogram Radiography Images Based on Convolutional Neural Network," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 7, no. 4, pp. 69-77, 2022.

[68] M. Rajesh, "Preprocessing and Skull Stripping of Brain Tumor Extraction from Magnetic Resonance Imaging Images Using Image Processing," Recent Trends in Intensive Computing, vol. 39, pp. 299 - 307, 2021.

[69] B. Perumal, J. Deny, S. Devi, & V. Muneeswaran, (2021, May). Region based Skull Eviction Techniques: An Experimental Review. In 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 629-634). IEEE.

[70] H.S.M. Chen, V. A. Kumar, J. M. Johnson, M. M. Chen, K. R. Noll, P. Hou, H. L. Liu, "Effect of brain normalization methods on the construction of functional connectomes from resting-state functional MRI in patients with gliomas," Magnetic resonance in medicine, vol. 86, no. 1, pp. 487-498, 2021.

[71] C. Salvatore, A. Cerasa, P. Battista, M. C. Gilardi, A. Quattrone, and I. Castiglioni, "Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach," Frontiers in neuroscience, vol. 9, p. 307, 2015.

Muhammad Irfan

Muhammad is pursuing PhD degree in Information Technology at the Western Sydney University. His research interest includes artificial intelligence, machine learning and deep learning.

Seyed Shahrestani

Seyed completed his PhD degree in Electrical and Information Engineering at the University of Sydney. He joined Western Sydney University in 1999, where he is currently a Senior Lecturer. He is also the head of the Networking, Security and Cloud Research (NSCR) group at the University.

Mahmoud Elkhodr

Mahmoud is Lecturer at Central Queensland University (CQU) Sydney Campus. His main teaching and research interests include Internet of Things (IoT), Computer Networking, Mobile technologies, health ICT, Security, and Privacy.

A Hybrid Multi-Person Fall Detection Scheme Based on Optimized YOLO and ST-GCN

Lei Liu[1], Yeguo Sun[2]*, Xianlei Ge[3]

[1] School of Computer Science, Huainan Normal University, Huainan (China)
[2] School of Finance and Mathematics, Huainan Normal University, Huainan (China)
[3] School of Electronic Engineering, Huainan Normal University, Huainan (China)

* Corresponding author: yeguosun@126.com

Abstract

Human falls are a serious health issue for elderly and disabled people living alone. Studies have shown that if fallers could be helped immediately after a fall, it would greatly reduce their risk of death and the percentage of them requiring long-term treatment. As a real-time automatic fall detection solution, vision-based human fall detection technology has received extensive attention from researchers. In this paper, a hybrid model based on YOLO and ST-GCN is proposed for multi-person fall detection application scenarios. The solution uses the ST-GCN model based on a graph convolutional network to detect the fall action, and enhances the model with YOLO for accurate and fast recognition of multi-person targets. Meanwhile, our scheme accelerates the model through optimization methods to meet the model's demand for lightweight and real-time performance. Finally, we conducted performance tests on the designed prototype system and using both publicly available single-person datasets and our own multi-person dataset. The experimental results show that under better environmental conditions, our model possesses high detection accuracy compared to state-of-the-art schemes, while it significantly outperforms other models in terms of inference speed. Therefore, this hybrid model based on YOLO and ST-GCN, as a preliminary attempt, provides a new solution idea for multi-person fall detection for the elderly.

Keywords

I. Introduction

THE elderly population is growing faster than any other age group, and as of October 2022, 10% of the world's total population will be over 65 years old [1]. Related studies predict that the total number of older adults will increase to 1.5 billion by the end of 2050 [2]. Older people's physical, cognitive, and motor skills decline with age. Falls are a significant challenge for them, and they can significantly reduce the life expectancy of older adults. Approximately 35% of people (65 years and older) fall once or more yearly [3]. In addition to old age, other factors such as environment, physical action, and cardiovascular disease can contribute to falls. It is a significant source of physical injuries, and these injuries usually require hospitalization for long-term treatment [4]. Each year, 37.3 million falls require medical care and 650,000 falls result in death [5]. In Fig.1, medical investigations have shown that timely treatment after a fall can reduce the risk of death by 80% and significantly improve the survival rate of older adults. Therefore, rapid detection of fall events is of great significance [6]. Accurate human action recognition methods and model optimization techniques are vital to achieving this goal.

However, there are also problems, such as significant differences in the structure and performance of each different model, low support for multi-person action recognition, and low real-time system performance [7]. This paper proposes a hybrid human fall detection framework based on YOLO and ST-GCN for multiple people for the elderly fall detection scenario in real time. Two optimization algorithms accelerate the model to improve real-time performance and accuracy of the model.

Fig. 1. Elder fall and timely treatment.

The primary advantages of the scheme are summarized as follows:

- We attempt to recognize human actions using skeletons and propose a hybrid model based on YOLO and ST-GCN to improve adaptation to multi-person fall detection scenarios.

- We accelerate the proposed hybrid model by using two model optimization algorithms to reduce the model size and improve the scheme's overall real-time performance.

- We construct our own fall detection test dataset for multi-person fall detection, and analyze the causes of miss and false detection in the fall detection model and provides references for subsequent research.

The rest of this paper is organized as follows: Section 2 reviews the research related to human fall detection and model optimization. Section 3 presents the proposed scheme's overall design framework and each component's functions, including the details of the hybrid model and the model optimization algorithm. Section 4 presents the construction method of the multi-person fall detection test dataset and gives the experimental protocol and results to verify the effectiveness and feasibility of the proposed scheme. Finally, the paper discusses the conclusions and directions for future work.

II. Related Work

Human fall detection is an independent human action recognition research direction, and researchers have proposed various methods for different technical characteristics. We focus on two issues: the design of a human fall detection model for multi-person and the optimization scheme of the model.

A. Human Fall Detection

In Fig.2, multi-person fall detection can be seen as an extension of single-person fall detection technology, and fall detection belongs to the human action recognition research field. Currently, three main fall detection methods exist 1) Environmental-device fall detection approaches. Detection is based on the environmental noise formed when the human body falls, such as sensing changes in object pressure and sound to detect falls [8]. This method has a high false alarm rate and cost, which is rarely used. 2) Wearable-sensor fall detection approaches. Falls are detected using accelerometers and gyro-scopes [9]. This method requires a long time to wear sensors, which not only affects the comfort of human life but also increases the burden on the body of the elderly. The false alarm rate is high in complex environments. 3) Computer-vision fall detection approaches [10]. It can be divided into two categories: the traditional machine vision method extracts fall features, has low hardware requirements but is susceptible to environmental factors such as background and light changes, and has poor robustness. The other category is artificial intelligence methods, which use camera image data to train and infer convolutional neural networks. This type of solution has the features of high recognition accuracy, no perception, and low cost.

In vision-based human fall detection schemes, human information from multiple modalities can be used as features of the model, such as appearance, depth, optical flow, and human skeleton [11]. Among them, the human skeleton node usually complements other class of modal features, which can convey important information and performs better in model accuracy and robustness [12]. Human skeletal node data mainly contains two dimensions of information, the temporal dimension and the spatial dimension [13]. In this case, the temporal dimension is information about the nature of the action being performed and how it is being performed. We review approaches to such modeling, most of which rely on RNN (Recurrent Netural Network) or convolutional neural network (CNN). WRNNs are neural
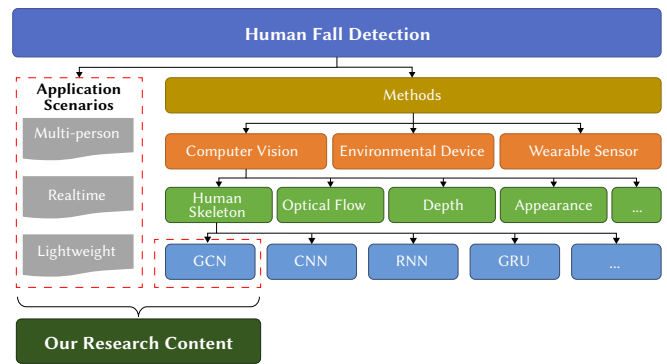
Fig. 2. Human fall detection method.

networks designed to process each time step of a time series one after the other, thus allowing the processing of variable length sequences. They maintain an internal state that captures the temporal context of the signal. RNNs are mostly based on long short term memory (LSTM) or gated recurrent units (GRU). Since the memory unit remembers values at arbitrary time intervals, it enables LSTMs to efficiently capture short-term and long-term temporal dependencies. In fact, LSTM have been widely used in human action-related problems, such as in action recognition [14], [15]. GRU, on the other hand, uses fewer parameters, and therefore less memory and lower computational cost [16], and therefore trains faster than LSTM. However, as shown by Weiss et al [17], LSTM outperforms GRU because it can easily perform unbounded counting, while GRU cannot. Thus, LSTM seems to be more accurate than GRU on longer sequences. In conclusion, the choice between LSTM and GRU depends on the data being processed and the application being considered.

On the other hand, spatial dimensionality, which is used to learn spatial correlation information in skeletal data [13], is modeled by three main categories: spatially- structured architecture (SSA), CNN, and graph convolutional network (GCN).

Among them, SSA relies on network architectures built around the human skeleton to help the model learn spatial correlations and allow the network to compute functions that essentially encode human skeletal features. These methods segment the skeleton into body parts and process the corresponding data in parallel network branches or hierarchical structures. In parallel approaches, the main distinction is usually related to the goal task, which determines the architecture of each branch [18]. Hierarchical approaches, on the other hand, model the human skeleton in layers, which can be either top-down or bottom-up [19], [20]. CNN is another type of architecture that relies on 2D convolution, i.e., in both the spatial and temporal domains. For this reason, the graph structure of the human skeleton is spread along the spatial dimension. CNN is particularly effective in learning spatial correlations in structurally regular data such as images. However, learning the spatio-temporal dynamics of human joints remains a challenge for CNNs because the graph structure of the human skeleton cannot be meaningfully flattened along a single dimension. The researchers have made some optimizations for this as well [21], [22].

Finally, as an extension of CNN, GCN have shown substantial performance advantages [23]. The GCN-based action recognition algorithm models human skeletal nodes as spatial-temporal relationships. It uses graph coarsening and partition design to enable GCN to process non-Euclidean data as efficiently as the human skeleton nodes and can achieve significant performance [24]. Generally, GCNs follow two major branches: Spectral GCN and Spatial GCN [25]. Spectral GCN implements graph con-volution for human skeleton-based action recognition by converting the graph from the

time domain to the frequency domain using the eigenvalues and eigenvectors of the graph Laplacian matrix [26] at the cost of extensive computation. Therefore, several measures are needed to reduce the computational cost of feature decomposition. In contrast, Spatial GCN has a lower computational cost and better performance, which has led to their more comprehensive application [27]. Therefore, most GCN-based approaches in human action recognition have focused on Spatial GCN. Human action is a continuous process, so time is crucial for representing human actions. Since Yan et al. proposed a spatial-temporal graph convolutional network (ST-GCN) in 2018, it has become a research hotspot [28]. Researchers have also proposed various improved versions of ST-GCN schemes. Peng [29] constructed a graph-based search space to explore the spatial-temporal connectivity relationships between nodes for action recognition. Shi [30] proposed that the adaptive learning graph structure is trained and updated along with the model parameters, which are better adapted to the action recognition task. Zhang [31] made the action recognition network more robust by introducing an attention mechanism. Cai [32] constructed a dual-stream model that combines the human pose skeleton and joint-centered lightweight information to capture the local delicate motions around each joint to improve the accuracy of action recognition.

In the multi-person fall detection scheme, [33], [34] used Long Short Term Memory (LSTM) for real-time multi-person fall detection and solved problems such as multi-person occlusion by using multiple cameras. Xu [35] improves the recognition accuracy for multiple users by using multiple trackers. Saturnino [36] proposes a hybrid fall detection model based on YOLO and SVM to improve the model's performance for multiple human targets detection.

Overall, the GCN-based schemes have higher inference accuracy in continuous actions sequences due to the inclusion of spatial and temporal feature information. However, compared to other schemes, GCN-based schemes generally have larger model sizes and do not have special treatment for multi-person scenarios, which is the issue we focus on.

B. Model Optimization

With the rise of Edge AI, more and more intelligent application scenarios occur at the edge end. In particular, some applications with high real-time requirements require the system to respond promptly in the production environment of the data [37]. How-ever, there is a significant contradiction between the vast scale of AI models and the constrained resources of edge devices [38]. While continuously improving the performance of Edge AI devices, accelerating the model inference rate through model optimization techniques is also the key to solving this problem [39], [40]. Among them, efficient network architecture design and model compression are typical approaches [41].

The modules are connected in the efficient network architecture design approach by creating a compact neural network structure and carefully designing the topology [42]. The goal is to achieve efficient deep learning models with acceptable accuracy while ensuring small model structure (low memory) and low computational complexity (high speed). For example, MobileNet and MobileNetV2 use deeply separable convolutional modules. In contrast, separable convolution with residuals/reverse residuals is its basic building block, significantly reducing the parameter size and achieving higher accuracy [43]. Similar examples of building blocks of efficient neural networks to reduce parameters and improve efficiency include ShuffleNetV1 and ShuffleNetV2 [44].

Unlike the efficient network architecture design approach, the model compression method aims to modify a given neural model to reduce its storage and computational costs [45]. In Fig. 3, the neural network-based model compression methods include pruning, quantization,

low-rank factorization, and knowledge distillation. Among them, pruning is one of the powerful techniques to remove unimportant components from the model [46]. Pruning is flexible and efficient in removing layers, neurons, connections, or channels, reducing the model by removing redundant parts. The purpose of quantization is to reduce the number of bits required for the model parameters to reduce the cost of storage and computation of the model parameters. Most processors use 32 or more bits to store the parameters of a deep model, which are stored in 16 or 8 bits after processing by quantization [47]. Knowledge distillation and low-rank factorization are also famous and influential in compressing depth models [48], [49].
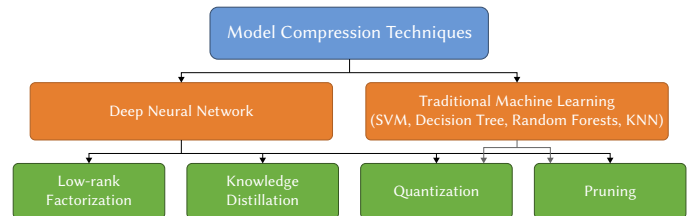
Fig. 3. Model compression method.

III. The Proposed Scheme

In Fig. 4, the traditional human action recognition scheme based on human skeletal nodes includes human posture estimation and human motion recognition. In the human pose estimation phase, the human skeletal node is extracted by human pose detection systems such as Microsoft Kinect and OpenPose and then handed over to the human action recognition model for processing. As mentioned earlier, ST-GCN is based on spatial-temporal information of human skeleton nodes and performs well in long action recognition, so our paper will use ST-GCN in the action recognition phase. ST-GCN generally uses OpenPose to extract skeleton nodes, but OpenPose usually extracts skeleton node information of multiple people from the global level at one time. This method has problems such as poor target recognition accuracy and poor interference resistance between skeleton nodes. In response, this paper optimizes the multiple human target detection process by adding the object detection model YOLO before the human pose estimation phase. The scheme starts with YOLO processing the original real-time video to generate multiple independent human detection boxes. Then each box is handed over to the human pose estimation model. Finally, the extracted human bones are processed by ST-GCN separately. The system's detection accuracy for multiple human targets can be improved by incorporating YOLO. In addition, by adopting a detection scheme based on small image blocks, the system dramatically reduces the computational load of the pose estimation model. Finally, this paper will accelerate the model inference rate through model optimization techniques.

A. Multi-Person Detection With YOLO

YOLO is an object detection model that uses a prediction method based on the whole frame [50]. After scanning an image only once, YOLO can detect all target information, including category and location, and performs well in object detection tasks. The latest version of the YOLO series is YOLOv7, which is currently the fastest object detection model [51]. Previous versions of YOLO have also differed significantly regarding network structure, parameter size, and model performance. Among them, the YOLOv5 version has a more balanced performance in all aspects and is widely used in various application scenarios. Unlike previous versions, YOLOv5 implements a series of network architectures, namely YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. These five models are similar in structure.
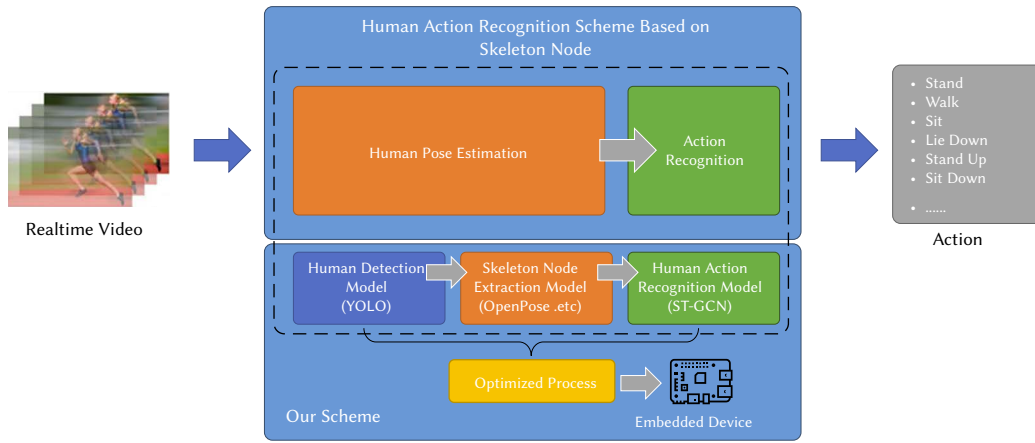
Fig. 4. The main framework of multi-person fall detection scheme.

TABLE I. Comparison of YOLOv5 Series Model

Methods	YOLOv5n	YOLOv5s	YOLOv5m	YOLOv5l	YOLOv5x
Depth multiple	0.33	0.33	0.67	1.00	1.33
Width multiple	0.25	0.50	0.75	1.00	1.25
C3-n (True)	1,2,3,1	1,2,3,1	2,4,6,2	3,6,9,3	4,8,12,4
C3-n (False)	1	1	2	3	4
Convolution kernels	16,32,64,128,256	32,64,128,256,512	48,96,192,384,768	64,128,256,512,1024	80,160,320,640,1280
Params (MB)	3.90	14.10	40.80	89.30	166.00
Speed (ms)	6	7	14	25	47
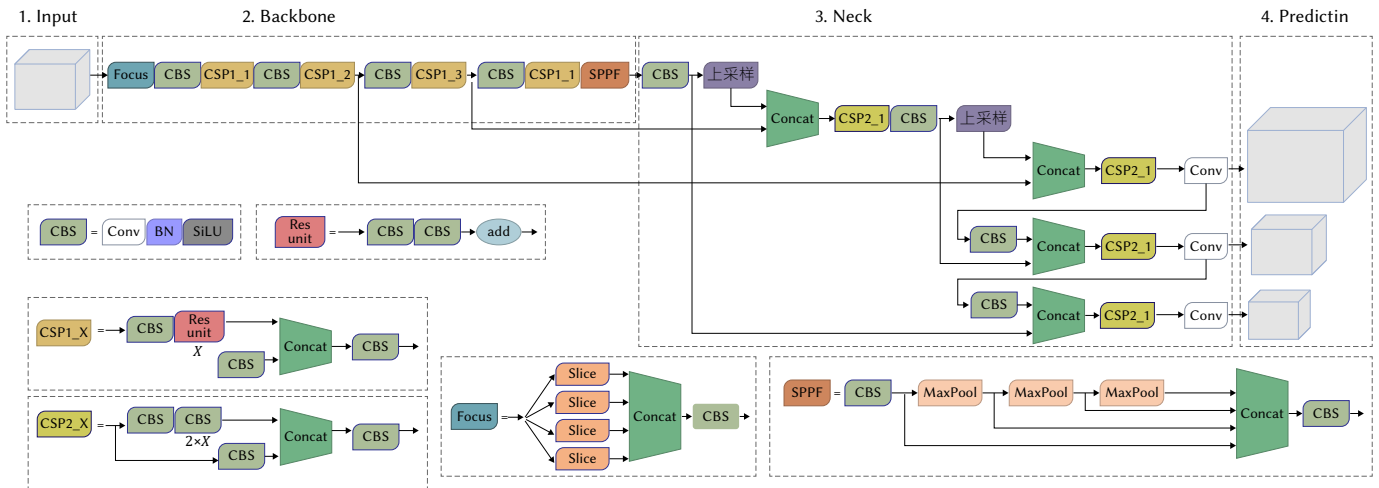
Fig. 5. Illustration of YOLOv5s's network.

The number of convolution kernels in the convolution process can be varied by changing the depth multiplier and the number of C3S in BottleneckC3 (Bottleneck CSP structure with 3 CBS modules) and the width multiplier. It allows for combinations between different network depths and widths to achieve a balance between accuracy and efficiency, as shown in Table I.

In the latest YOLOv5 series, the model size is 3.87MB for YOLOv5n and 14.1MB for YOLOv5s. They are low-cost target detection models suitable for deployment on mainstream mobile or edge devices. In the COCO data set, YOLOv5n was used for object detection, 26.4% of the images had missing person detection, and the time of each image was 6ms on average. YOLOv5s was used for target detection, and 13.8% of the images were missing people detection, and the average time of each image was 7ms. To ensure fall detection accuracy, YOLOv5s is chosen as the base model in this paper for optimization and improvement.

Fig. 5 shows the network framework of YOLOv5s-6.0, which consists of four parts: 1) Convolutional network-based Backbone network, which mainly extracts image feature information. 2) Head detection head, main prediction object box, and prediction object category. 3) The Neck layer between the trunk network and the detection head. 4) The prediction layer outputs the detection results and predicts the object detection frame and label category.

In Fig. 6, this paper obtains video data from the camera and processes each image frame. After YOLO processing, the system will detect multiple human targets. These objects will be given to the human pose estimation model as independent image blocks to extract human skeleton nodes. Since the entire image does not need to be searched during the skeleton node extraction phase, it can be directly based on the independent image block containing the human object, which will significantly improve the inference speed and accuracy of the model.
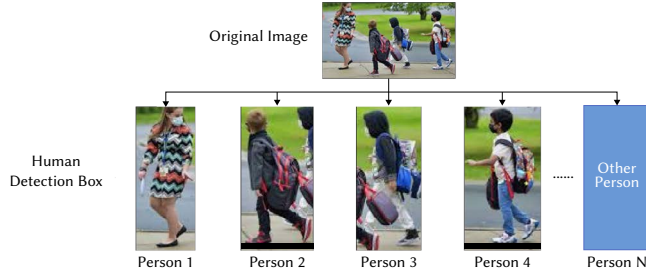
Fig. 6. Human detection box in YOLO.

B. Multi-Person Fall Detection With ST-GCN

ST-GCN pioneered the application of GCN in human action recognition based on skeletal nodes. In Fig. 7, the input to ST-GCN is a joint coordinate vector of a graph node that is obtained based on one graph node and two graph edges. The graph nodes are the skeletal nodes, and the graph edges are the spatial and temporal edges of the skeletal nodes. One of these is the spatial edge between the different skeletal nodes, representing the skeletal constraint information of the human action. The other category is the temporal edges, which are connected between the same skeletal nodes at different moments and represent the temporal constraint information of the human action. These data extract high-level features through the spatial-temporal graph convolution operation. Then the corresponding action classification is obtained as the output using the SoftMax classifier. ST-GCN integrates the temporal and spatial information of skeletal nodes in human actions and performs well in long-action recognition. In this paper, we use ST-GCN for human action recognition.
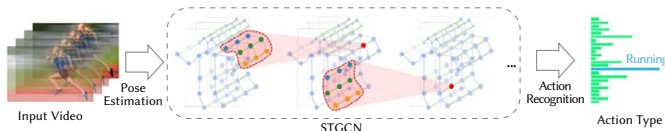
Fig. 7. Illustration of ST-GCN.

The principle and processing flow of the multi-person fall detection method designed in this paper is similar to that of the single-person scheme. In Fig. 8, YOLO identifies multiple human targets from video images and sends them to the human pose estimation model for processing in the form of a human image detection frame. The human posture estimation model extracts bone nodes from each image detection box and further generates a continuous bone image sequence, and ST-GCN will use this for action recognition. Such a scheme that separates the object detection part from the action recognition makes our scheme able to improve the recognition accuracy and speed of the whole system for the actions of multiple people just by adopting the object detection model similar to YOLO.

ST-GCN is a human action recognition model based on skeleton nodes, which can be extracted using OpenPose. OpenPose is a convolutional neural network based on Caffe, a real-time 2D multi-person pose estimation model developed by Carnegie Mellon University (CMU) [52]. It is a bottom-up human pose estimation model which can realize the pose estimation of human movement, facial expression, and
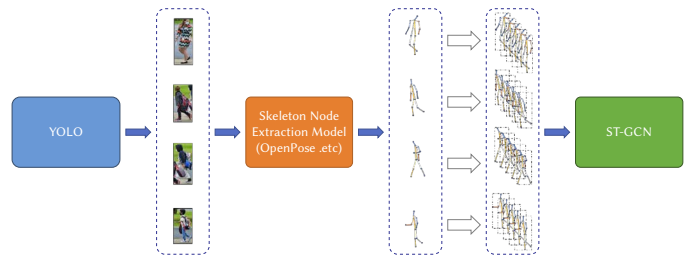
Fig. 8. The main framework of our scheme.

finger movement. It is suitable for single and multi-person scenarios, with excellent recognition effects and fast recognition speed. Fig. 9(a) shows that OpenPose commonly uses the skeletal model containing 25 nodes. Dot 0 represents the nose; Dot 15 to 18 represents the left and right eyes and ears; Dot 1 represents the neck; Dot 2 to 7 represents the left and right shoulders, elbows, and wrists; Dot 8 represents the center of the buttocks, dot 9 to 14 represents the left and right hips, knees and ankles; Dot 19 to 24 represents the left and right nodes of the feet, toes, and heels. To improve the processing efficiency of the model, we simplified the skeletal node structure, deleting one hip center node, 4 facial nodes, and 6 feet nodes. Finally, our skeletal model contains only 14 nodes, as shown in Fig. 9(b).
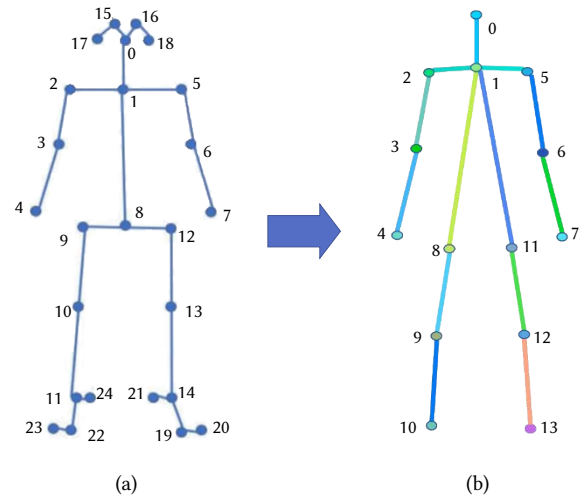
Fig. 9. Illustration of skeleton node structure.

Fig. 10 shows the simplified human skeleton nodes extracted by OpenPose. It can be seen that this method can accurately identify bone nodes in different scenarios. Both (a) and (b) are RGB images, and (c) is the depth image. Among them, the illumination condition of (a) is poor, and the illumination condition of (b) is better. After each frame, OpenPose will generate a 3*14 skeleton node matrix M. The M3*14 stores 14 skeleton nodes, and each node data contains a set of (X, Y) image coordinate information and the confidence Score of the node. The higher the value of the Score, the higher the accuracy of the predicted skeleton node.
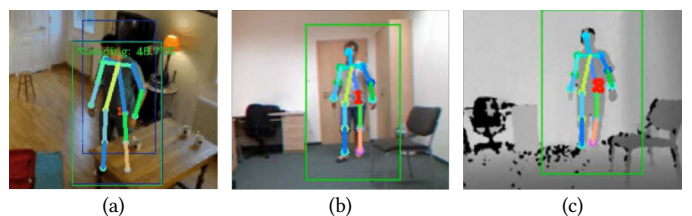
Fig. 10. Skeleton extraction in different scenes.

TABLE II. Comparison of Fall Detection Datasets

Dataset	Type	Number of samples	Remark
MCFD	Video	Total 24 videos 1) 14 fall videos 2) 10 non-fall videos	1) Each video contains eight videos from different perspectives
Le2i FDD	Video	Total 191 videos, totaling 75911 frames 1) 132 fall videos, totaling 43745 frames 2) 59 non-fall videos, totaling 32166 frames	1) 1 overhead camera angle 2) Coffee-Room and Home scenes are labeled 3) Lecture-Room and Office scenes are unlabeled
FDD	RGB Image Depth Image	Total 22636 images 1) 4212 fall images 2) 18424 non-fall images	1) 8 overhead camera angles 2) 5 different rooms
URFD	RGB Image Depth Image Video	Total 70 videos, totaling 14539 frames 1) 30 fall data sets, totaling5990 frames 2) 40 daily activity data sets, totaling 8549 frames	1) 2 camera angles are parallel to the ground and directly above the ceiling 2) 1 camera angle records average activity data, and the position is parallel to the floor

There are many kinds of fall detection datasets based on RGB images [53]. Standard lightweight datasets include the Multiple Cameras Fall Dataset (MCFD), Le2i Fall Detection Dataset (Le2i FDD), and the University of Rzeszow Fall Detection (URFD) Dataset. The details are shown in Table II. MCFD [54] contains 24 action sequences recorded by 8 cameras from different angles. The same subject performed the fall action and Activities of Daily Living (ADL), recording ten action types. Le2i FDD [55] used a single RGB camera, and 9 subjects performed 3 types of fall actions and six different ADLs. The videos were captured in 4 different environments (home, coffee room, office, and lecture hall). Actions are carried out in various factors such as light, clothing, the color of the dress, texture of clothing, shadows, reflections, camera view, etc. FDD [56] was recorded from 8 different viewpoints in 5 rooms. The study had five participants, including 2 males and 3 females. The actions performed by the participants included standing, sitting, lying down, bending over, and crawling, which were recorded at a rate of 30 images per second. URFD [57] was produced by the Interdisciplinary Center for Computational Modeling at Rzeszow University. The video sample contains 70 action sequences recorded at 30 frames per second. The dataset recorded falls and ADL, such as standing, bending, and lying down. The environment has adequate lighting. Although these datasets only contain single-person samples, since the fall detection model for multiple people designed in this paper is based on the single-person action recognition method, it can be used as a training dataset for ST-GCN.

In Fig. 11, the positions of the cameras of the dataset can be classified into three types depending on the application scenario. Among them, the height of the camera position in 10(a) is about 45 degrees of the elevation angle of the human line of sight. The device's height in 10(b) is about the waist of the human body. The equipment of 10(c) is located at the top of the ceiling. Generally, scheme (a) is more commonly used [58], [59]. Still, to avoid the influence of the impression model on action recognition due to the camera position factor, the samples of the training data set in this paper will be obtained from FDD and URFD [60]. We augment the original dataset with examples from these two datasets using image data augmentation methods such as symmetry inversion, motion blur, brightness change, and image rotation. After expansion, FDD contains 1084 groups, URFD has 847 groups, the sample resolution is 640*480.

We mix the two augmented datasets, on the one hand, to ensure a sufficient number of samples and, on the other hand, to increase sample diversity and avoid overfitting. In Table III, we extract a certain amount of fall action and non-fall action training samples from each dataset to form the Mix-Dataset of this paper for model training.

(a)   (b)   (c)
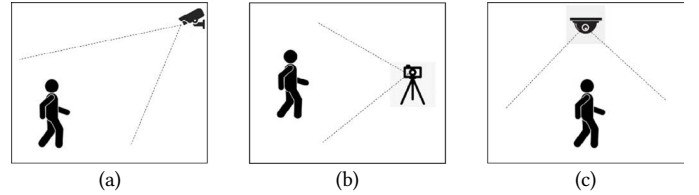
Fig. 11. Camera location.

TABLE III. The Dataset Composition

Dataset	Fall Action (Groups)	Non-Fall Action (Groups)
FDD	550	300
URFD	380	270
Mix Dataset	930	570

The Mix Dataset samples are all images that can be directly used to train CNN-based and RNN-based fall detection models. However, the GCN-based scheme processes the action sequence data based on time series, and the images need to be further processed into the kinetics-skeleton format required by ST-GCN. The kinetics-skeleton converts a sequence of skeletal actions into a list $V_{List} = \{V_1, V_2, ...V_n\}$. Each V_k represents one image, which consists of three parts. $V_k = \{frame\_index=k, skeleton \{pose[p_1, p_2, ...p_m], score[s_1, s_2, ...s_p]\}\}$. Fig. 12(a) shows that the *frame_index* represents the frame number in the action sequence. The *pose* represents the point information in the transformed graph vector of the current skeleton node, which indicates the state of the skeletal node. The *score* represents the edge information in the transformed graph vector of the current skeleton node and indicates the state between skeleton nodes. Fig. 12(b) is the label information of this skeleton sample. In our paper, we set $n = 150$, $m = 35$, and $n = 17$. It is expressed as 300 frames per training sample, about 10 seconds per video calculated with 30 frame/s.

(a)   (b)

Fig. 12. Kinetics-skeleton node format.

C. Model Optimization

To further improve the real-time performance of the proposed scheme, we use two methods to optimize the model. In particular, we optimize the YOLO by enhancing the model network structure and use the AI acceleration framework-TensorRT to reoptimize the YOLO and ST-GCN.

1. Optimized Design of YOLO

The CSPDarkNet53 backbone network used in YOLOv5s is a Cross Stage Partial Network (CSPNet) introduced in Darknet53 [23] to extract sufficient depth feature information. To enable the YOLO with a more robust feature extraction capability, we used MobileNetV3 to attempt to achieve a coordinated balance of lightweight, accuracy, and efficiency. MobileNetV3 is a light backbone network that performs better on the edge and mobile side [61].

MobileNet (i.e., MobileNetV1) is a lightweight CNN, which is more suitable for deployment on the edge devices. It can use Depthwise Separable Convolution (DSC) to vary the computation of convolution to reduce the number of network parameters to balance accuracy and efficiency. MobileNetV2 adds two new modules: Reverse Residuals (IR) and Linear Bottlenecks (LB). The IR module can make the model have better feature transmission capability and deeper network layer. Meanwhile, MobileNetV2 uses LB module instead of the non-linear module, thus reducing the loss of the model on low-level features. MobileNetV3, released in 2019, combines some of the structures from V1 and V2 and removes the more computationally expensive network layer from the V2 architecture. It achieves low resource consumption while guaranteeing almost no loss of accuracy by introducing the lightweight attention structure of Squeeze and Excitation Networks (SE-Net) [62].

In Fig. 13, we replace the backbone of YOLOv5s with the feature extraction network of MobileNetV3. In the YOLOv5s, three different sizes of feature maps can be obtained after three down-sampling, and then feature fusion is performed. The locations of the three down-samplings are identified with a,b,c. We choose the output of the last three down-samplings of the MobileNetV3 feature extraction network as an alternative. Specifically, the feature extraction network of MobileNet3 contains 13 sampling modules Module [0-12]. Among them, Module[4] is the penultimate third down-sampling, so Module[0-3] is used as MobileNet1, whose down-sampling position is identified with 1; Module[9] is the penultimate third down-sampling, so Module[4-8] is used as MobileNet2, whose down-sampling position is identified with 2; and, finally, Module[9-12] as MobileNet3, whose down-sampling position is identified with 3. The results of these three down-samplings will be used for subsequent processing in YOLOv5s.
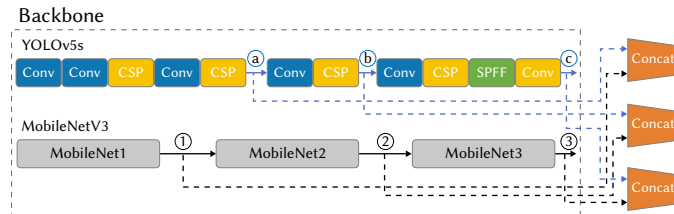
Fig. 13. Depthwise separable convolution framework.

DSC consists of Depthwise Convolution (DW) and Pointwise Convolution (PW) [63], as shown in Fig. 14, and the parameters and computational effort of DSC are significantly reduced compared to traditional convolution. A comparison of the computational effort between the two is shown in (1). W1 and W2 are the computational costs of DSC and the standard conventional convolution, respectively. The size of the convolution kernel for MobileNetV3 feature extraction is mainly 5×5. Therefore, the computational cost of DSC is about 1/25 of the traditional conventional convolution.
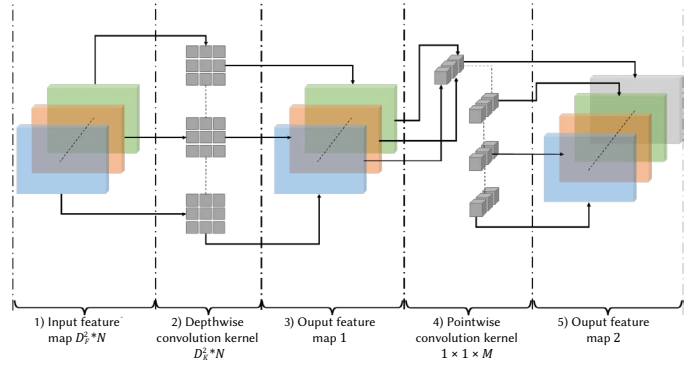
Fig. 14. Depthwise separable convolution framework.

$$\frac{W_1}{W_2} = \frac{D_k^2 \times M \times D_F^2 + M \times N \times D_F^2}{D_K^2 \times M \times N \times D_F^2} = \frac{1}{N} + \frac{1}{D_K^2} \tag{1}$$

The Fig. 15(a) is the residual network structure, and Fig. 15 (b) is the reverse residual network structure. Reverse residual networks use point convolution to expand the number of channels, then deep convolution in higher layers, and finally, use point convolution to shrink the channels. Reverse residual networks improve the gradient propagation of features with the help of residual connections, making the network layer deeper. The network uses minor input and output dimensions, significantly reducing the network's computational consumption and parameter size. In addition, the reverse residual network is CPU and memory efficient for inference and enables the construction of flexible mobile-side models, making it suitable for edge side applications.
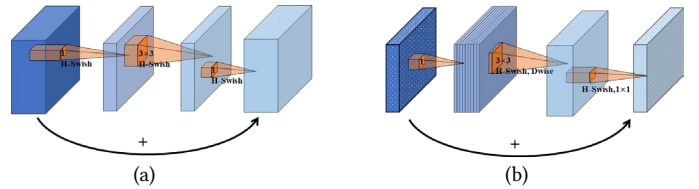
Fig. 15. Residual network and reverse residual network.

MobileNet includes two hyperparameters, α, and β. Where α represents the width factor, which can be adjusted to α times the original convolution kernel by adjusting the number of convolution kernels, and β represents the control input image size. In this paper, the amount of computation after adjusting α using DSC in (2).

$$W = D_K^2 \times \alpha M \times D_F^2 + \alpha M \times \alpha N \times D_F^2 \tag{2}$$

By adjusting the α, the calculation effort and model volume can be directly reduced to $1/\alpha^2$, which significantly reduces the model's number of parameters and computational effort with minimal loss of accuracy. We set α=0.5, and the optimized model is YOLOv5s-opt. In this paper, SSD, Faster-RCNN, YOLOv4, YOLOv5s, and Yolov5s-opt are tested on the COCO dataset, and the results are shown in Table IV. It can be seen that the optimized YOLOv5s-opt reduces the parameter size by 50% and improves the frame rate by 15% compared with YOLOv5s.

2. Optimized Design of ST-GCN

In this paper, TensorRT will be used to optimize ST-GCN. It is an AI optimization and deployment framework designed by NVIDIA for GPU [64]. In Fig. 16, it is both an inference optimization engine and a runtime execution engine. It provides optimal support for the model's inference at the graphics optimization, operator optimization, memory optimization, and Int8 calibration levels. Specifically, it benefits from the fact that after training the neural network, TensorRT can compress,

TABLE IV. Performance Comparison of Object Detection Models

Model	Params (MB)	FPS (frame/s)
SSD	100	67
Faster-RCNN (ResNet50)	109	42.9
YOLOv4 (CSPDarkNet53)	24.5	31.5
YOLOv5s (CSPDarkNet53)	14.1	61.5
YOLOv5s-opt (ours)	7.2	70.2

optimize, and deploy the network at runtime without the overhead of a framework. It can also improve the latency, throughput, and throughput of the network through combining layers, kernel optimization selection, as well as performing optimization and conversion to optimal matrix math methods based on a specified precision.
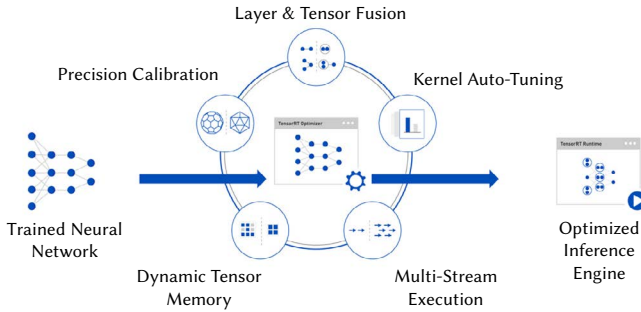
Fig. 16. Model optimization processing flow of TensorRT.

The operation of TensorRT consists of two main phases, Build, and Deployment. The Build phase involves the conversion of the model from another model form to a TRT form. During the model conversion, the inter-layer fusion and accuracy calibration of the optimization mentioned above is completed. The output of this step is an optimized TRT model for the specific GPU platform and network model, serialized to disk or memory in the form of a plan file. The plan file from the previous step is first deserialized, a Runtime Engine is created, and the inference task can then be executed. The YOLO and ST-GCN are built in the PyTorch framework and converted into TRT models using ONNX intermediate conversions [65], as shown in Fig. 17. After optimization, the model can reduce the model's parameters by 16%.
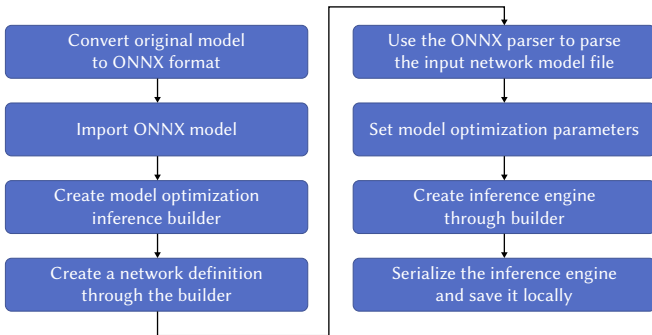
Fig. 17. The optimization process of YOLO and ST-GCN.

IV. Experiments and Discussions

Experimental hardware: CPU: 11th Gen Intel Core (TM) i7-11700 @2.50 GHz. Memory: 16GB; GPU: NVIDIA GeForce GTX 1080 Ti. GPU Acceleration Library: CUDA 11.0.3, CUDNN 8.2.1. OS: Windows 10 (64-bit). Software tools: OpenCV 4.1.1, Pytorch 1.7.0, TensorRT 7.1.3.

A. Dataset

The fall test datasets in this paper are divided into two types: single-person and multi-person. Among them, the original single-person dataset was selected from the publicly available dataset Le2i FDD with 155 human fall videos, which include 95 ADL videos and 60 fall videos. Each fall action had a complete fall process containing the other action to fall.

In Fig. 18, the scenes include 3 types: home, office, and pantry. The home scene (a) is a living room scene, including sofas, dining tables, stairs, chairs, table lamps, and other accessories, and contains a variety of lighting conditions. The office scene (b) includes tables and chairs with a more regular and balanced light distribution. The pantry scene (c) includes sofas, tables, tea sets, etc. The light is more frequent and evenly distributed. The video resolution is 320×240.

Fig. 18. Original single-person fall detection test dataset.

To experiment more effectively, we created our multi-person fall dataset (MPFDD), which has two scenarios for 2 to 5 persons, respectively. The original dataset consisted of 220 videos divided into indoor and outdoor scenes. It consists of 80 ADL videos and 140 fall videos. The indoor scene is the action room scene, which includes chairs, tables, computers, and other accessories. The outdoor scene is open, with tables and chairs as the main accessories. Both scenes of 2-person and 3-person have good lighting conditions, but it is no need in 4-person and 5-person. In addition, the 2-person scene includes 20 ADL videos, 10 fall videos with 1-person, and 10 fall videos with 2-person. The 3-person scene have 20 ADL videos, 10 fall videos with 1-person, 10 fall videos with 2-person, and 10 fall videos with 3-person. The 4-person scene have 20 ADL videos, 10 fall videos with 1-person, 10 fall videos with 2-person, 10 fall videos with 3-person, and 10 fall videos with 4-person. The 5-person scenes have 20 ADL videos, 10 fall videos with 1-person, 10 fall videos with 2-person, and 10 fall videos with 3-person, 10 fall videos with 4-person, 10 fall videos with 5-person. The video resolution is 1060×510. In Fig. 19, we show some of the videos in the original MPFDD.

Similar to the train dataset, the original single-person dataset and multi-person dataset were expanded using image data augmentation techniques. The main data augmentation algorithms we use include symmetric flipping, adding Gaussian noise, motion blur and brightness contrast transformation.

In Fig. 20, (a) represents the original video, and (b)-(e) represent the effects after being processed by the above four data augmentation algorithms, respectively. We do not process every original video with all four image data augmentation algorithms. In particular, we do not use the brightness contrast transform algorithm for videos that are not very bright. In the end, we get the expanded dataset, which consisted of a total of 1044 videos, which included 413 ADL videos and 631 fall videos.

B. Evaluation Metric

In classification tasks, the confusion matrix is often used to show the results predicted by the model. In this case, the actual classes are represented by the columns of the matrix, while the rows indicate the predicted classes [66]. For each class, the matrix displays true positive
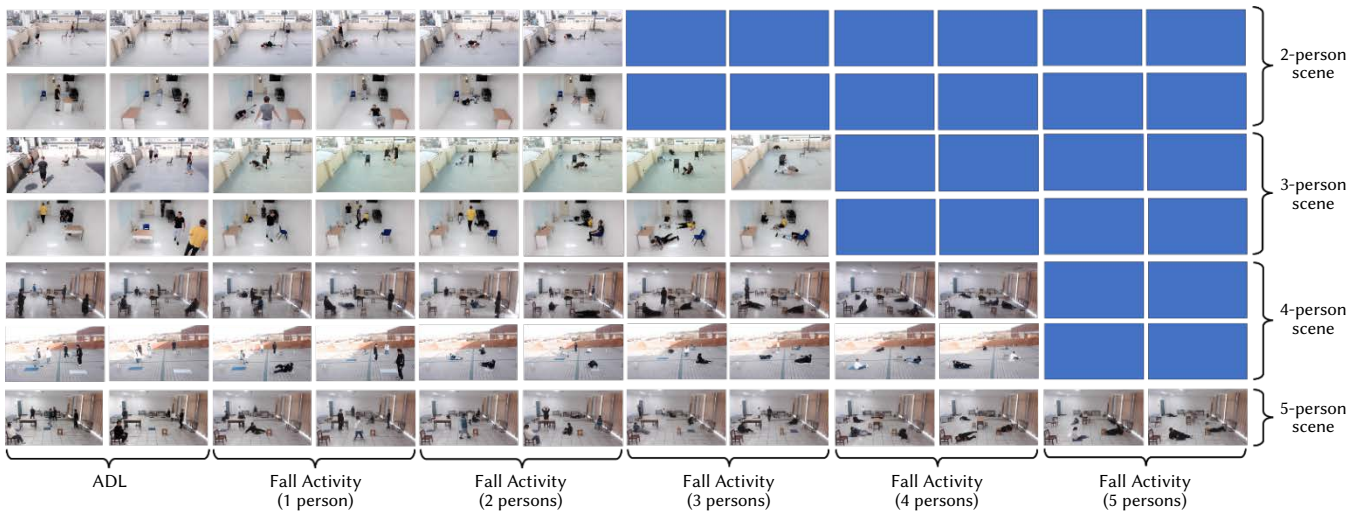
Fig. 19. Original Multi-person fall detection test dataset (MPFDD).
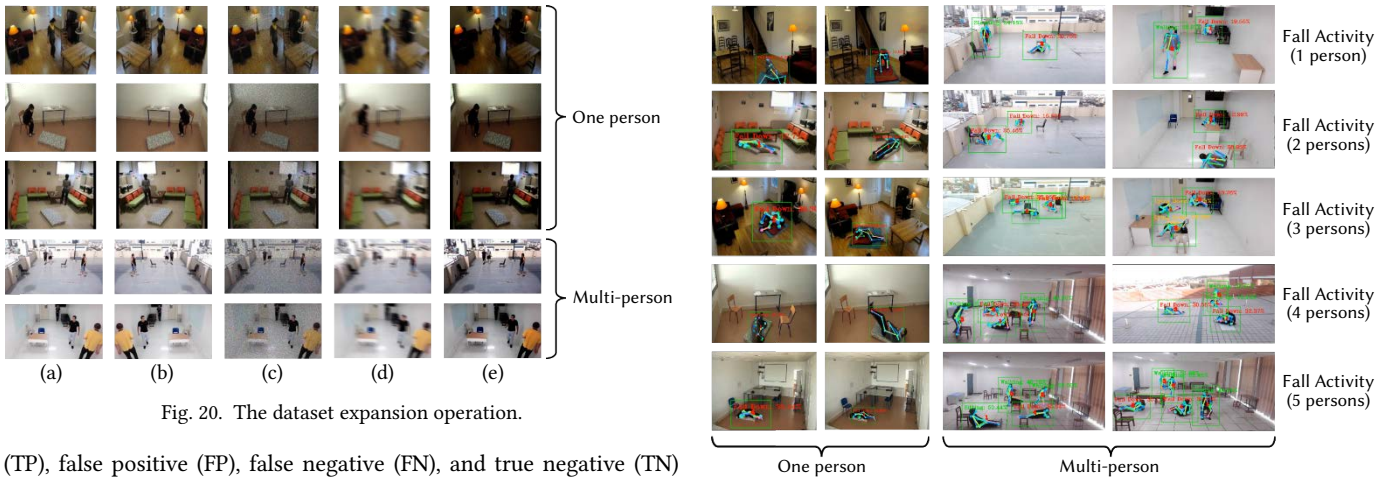
Fig. 20. The dataset expansion operation.

Fig. 21. Multi-person fall detection.

(TP), false positive (FP), false negative (FN), and true negative (TN) values. Where TP is the number of fall targets correctly detected. FP is the number of targets that were falsely detected by falls. FN is the number of samples where falls were not detected. TN is the number where no falls were correctly detected. The confusion matrix can calculate many model evaluation metrics, including precision, recall, accuracy, and F1 score [43], as shown in (3), (4), (5) and (6). In addition, there is the criterion FPS which measures the real-time performance of the model. It indicates the number of images processed per second. In this paper, the models are evaluated according to the above criteria.

$$Precision = \frac{T_P}{T_P + F_P} \times 100\% \tag{3}$$

$$Recall = \frac{T_P}{T_P + F_N} \times 100\% \tag{4}$$

$$Accuracy = \frac{T_P + T_N}{T_P + F_N + T_N + F_P} \times 100\% \tag{5}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

C. Results and Analysis

Fig. 21 shows the results of the proposed scheme for multi-person fall detection. There are various fall detection schemes, mainly including CNN-based method, RNN-based method, GCN-based method, and some other types of methods. We select two typical methods in each type of scheme to compare with our proposed scheme. The comparison results are shown in Table V. Among them, Carlier [67] and Zhang [68] use CNN-based methods. Kareem [69] and Yadav [70] use RNN-based
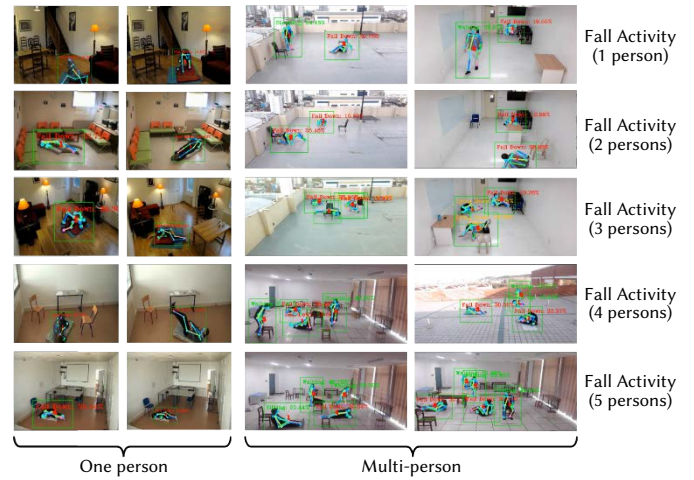
methods. Zheng [71] and Lee [72] use GCN-based methods. The above methods are all estimate human skeleton nodes as features. In addition, Maheswari [73] and Kiran [74] are representatives of two other types of schemes. Maheswari uses the human body as the entire modeling object and the settlement results of HOG and SRMAR as features for detection. Kiran extracts human behavioral features at different stages through a combination of multiple CNNs, and introduces SVM models for inference of the results. We also use an optimization scheme based on ST-GCN, so it is also incorporated. Miss detection represents the number of samples in the fall video that fails to detect fall action, and false detection represents the number of samples in the fall test video that detect non-fall action as a fall action.

In Table V and Fig. 22, we can see that: 1) the CNN-based scheme is low in terms of accuracy, while the RNN- and GCN-based schemes are relatively high, but the former has an advantage in terms of frame rate. This is because the CNN-based scheme uses prediction based on the spatial information of a single skeletal node. In contrast, the latter two are based on processing a sequence of skeletal nodes with additional temporal information. As a result, the former is less computationally intensive and faster, while the latter has higher detection accuracy for predicting long movements. 2) Our scheme performs better in all metrics compared to other schemes. Our scheme has a less obvious advantage in terms of accuracy since various optimized versions of pose estimation models have been proposed in recent years to improve the accuracy of extracting skeletal nodes. The scheme represented by

TABLE V. Performance Comparison of Different Algorithms

Paper	Method	Miss Detection	False Detection	Precision (%)	Recall (%)	Accuracy (%)
[67]	Optical-Flow based CNN	64	25	75.9	74.3	74.7
[68]	VARNN+CNN	41	29	78.0	77.2	78.0
[69]	Optimized RNN	40	19	81.8	80.1	81.4
[70]	ConvLSTM	35	14	82.4	81.6	82.2
[71]	Optimized GCN	43	20	82.9	80.4	81.1
[72]	2s-AGCN	19	14	86.5	85.7	86.2
[73]	HOG+SRMAR	44	29	79.5	78.0	79.2
[74]	Multi-CNN+SVM	57	45	73.5	71.2	72.0
--	ST-GCN	36	31	81.2	79.8	80.8
--	Ours	28	17	84.3	82.7	83.8

[72] uses a two-stream algorithm (2s-AGCN) that can model both first-order and second-order information, significantly improving the recognition accuracy. In comparison, the overall testing performance of both [73] and [74] can reach over 70%, and the inference speed of these two methods is also higher, mainly due to the smaller scale of the model. The model structure used in [74] is relatively simple, and the ability of CNN to extract deep-seated human motion features is limited, and it cannot integrate temporal and spatial information of motion, so its inference accuracy is low. In summary, our method can achieve good performance and speed trade off.

The second and third columns of Table V show that all schemes in the experiment have miss detection and false detection. We analyzed the reasons for these situations. In the experiment, we found that the causes of miss detection and false detection in the multi-person scene are similar to those in the single-person scene. We mainly use the video display effect of single-person scenes to simplify the analysis content and process. The comparison model includes [67], [69], [71], ST-GCN, and ours.
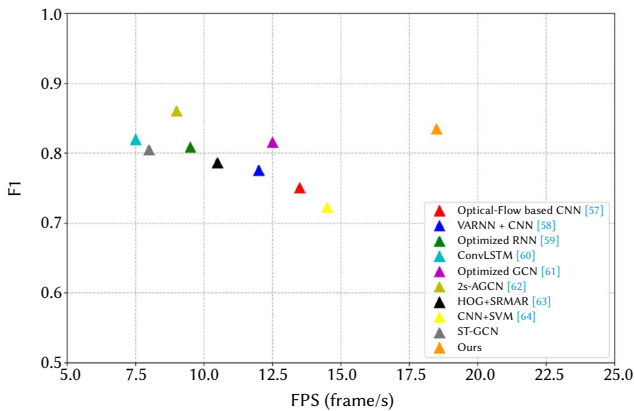
Fig. 22. Multi-person fall detection.

In Fig. 23, the main reasons for miss and false detection can be summarized into four categories, and the red "□" represents the human position in the video. 1) The part of the human body is close to the edge of the image, resulting in the partial loss of the part of the human body image, which is not conducive to the extraction and discrimination of human skeleton features. In (a), most of the head and upper body of the tester are outside the image range, and all test models fail to detect them. 2) Due to the uneven distribution of light in the scene, the gradient features of the human body are not apparent, which affects the extraction of skeletal nodes. In (b), the clothes of the tester and

the ambient background were red, and the ambient luminance was insufficient. Although all test models could detect the human target, the extracted skeletal states deviated severely from reality. 3) Skeletal feature extraction failed due to partial or complete occlusion of the human body parts. This is similar to the first case, where the occlusion can be a stationary accessory or a moving person in the scene. In (c), the test person's head and the right half of the body are obscured by the table, and all test models fail to detect. 4) Due to the perspective effect of the camera, when the test tester falls in a direction parallel to the direction of the camera's vision and when the head is further away from the camera than the feet, the tester's skeletal state is projected onto the image in a state of action similar to ADL, which leads to miss detection. In (d), the tester is in a semi-slumped state parallel to the camera's vision direction, with the skeleton showing a state similar to sitting and bending, and all test models fail to detect it.

In the multi-person scene, it is more evidence that some human targets fail to be detected due to mutual occlusion between human bodies. As shown in Fig. 24, the red "□" represents undetected human targets. In short, the main reason for the miss and false detection is the inaccurate extraction of human skeleton node features caused by various environmental factors, which will provide a clear research direction for our further work.

V. Conclusions and Future Scope

To improve the quality of life of the elderly, we propose a fall detection scheme based on human skeleton nodes. This scheme is a hybrid model based on YOLO and ST-GCN, which can support multiple fall detection. We also use model optimization technology to accelerate the proposed model, further reduce the scale of the model and improve the inference speed. The experimental results show that in a good testing environment, this scheme has high detection accuracy and obvious real-time advantages. Therefore, as an attempt at a hybrid model for multi person fall detection, this scheme has some reference value for subsequent research. Our scheme also has problems because the detection cannot be recognized due to uneven light distribution, blocked human body parts, and unique fall direction. To improve the detection accuracy of the proposed model, the following aspects will be studied in future work:

- Study of light adaptive compensation algorithms. Incorporate it into a fall detection system to increase the system's resistance to light changes.
- Study of multi-camera detection methods. Try using multiple cameras for fall detection from various angles to solve the occlusion problem.
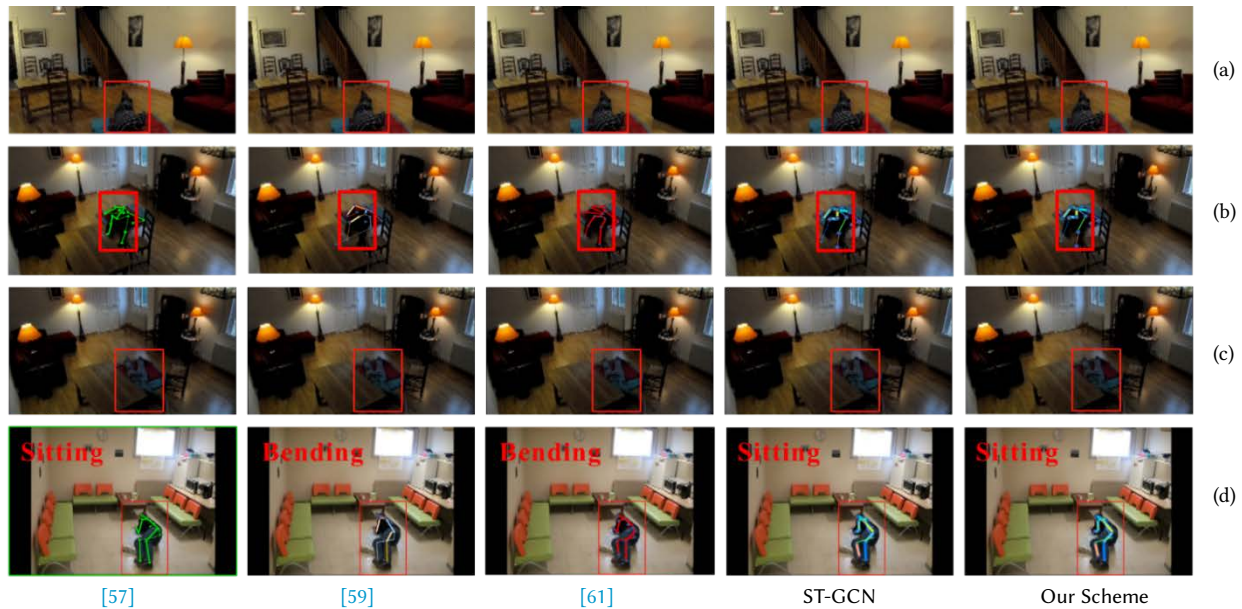
Fig. 23. Miss detection and false detection of different algorithms.
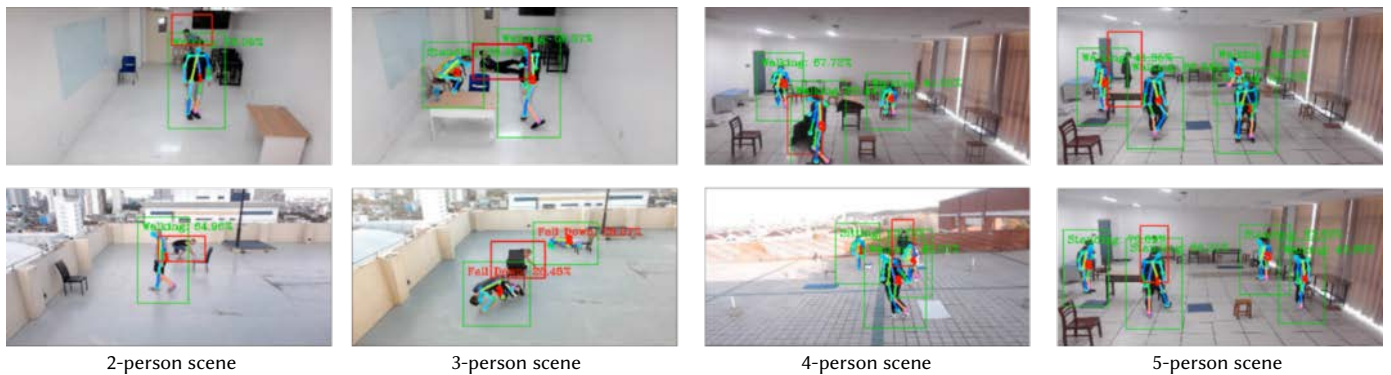
Fig. 24. Miss detection in multi-person scene.

- As there is no publicly available public dataset dedicated to the abnormal action of the elderly, such data is missing from our test dataset MPFDD. In future work, we will gradually collect data on the abnormal action of the elderly through cooperation with relevant medical institutions and elderly care institutions. We will also improve the fall experiment by adding more human actions and test scenarios to optimize the proposed scheme's shortcomings further.

References

[1] Ageing and health, World Health Organization., 2022. Accessed: June. 8, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health.

[2] World Population Ageing: 1950–2050, Global Action on Aging., NY, USA, 2002. Accessed: June. 8, 2024. [Online]. Available: http://globalag.igc.org/ruralaging/world/ageingo.htm.

[3] S. Usmani, A. Saboor, M. Haris, M. A. Khan, and H. Park, "Latest Research Trends in Fall Detection and Prevention Using Machine Learning: A Systematic Review," *Sensors*, vol. 21, no. 15, pp. 5134-5156, 2021, doi: 10.3390/s21155134.

[4] S.-H. Jung, J.-M. Hwang, and C.-H. Kim, "Inversion Table Fall Injury, the Phantom Menace: Three Case Reports on Cervical Spinal Cord Injury," *Healthcare*, vol. 9, no. 5, pp. 492-500, 2021, doi: 10.3390/healthcare9050492.

[5] Falls, World Health Organization., 2021. Accessed: June. 8, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/falls.

[6] H. Ramirez, S. A. Velastín, I. Meza, E. Fabregas, D. Makris, and G. Farias, "Fall Detection and Activity Recognition Using Human Skeleton Features," *IEEE Access*, vol. 9, pp. 33532-33542, 2021, doi: 10.1109/ACCESS.2021.3061626.

[7] H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastín, "Video-based Human Action Recognition using Deep Learning: A Review," *ArXiv*, vol. abs/2208.03775, pp. 1-25, 2022, doi: 10.48550/arXiv.2208.03775.

[8] X. Li, J. Li, J. Lai, Z. Zheng, W. Jia, and B. Liu, "A Heterogeneous Ensemble Learning-Based Acoustic Fall Detection Method for Elderly People in Indoor Environment," *Artificial Intelligence in HCI: First International Conference, AI-HCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22.* Springer International Publishing, 2020, pp. 369-383., doi: 10.1007/978-3-030-50334-5_25.

[9] A. Sucerquia, J. D. López, and J. F. Vargas-Bonilla, "Real-Life/Real-Time Elderly Fall Detection with a Triaxial Accelerometer," *Sensors*, vol. 18, no. 4, pp. 1101-1118, 2018, doi: 10.3390/s18041101.

[10] J. Gutiérrez, V. Rodríguez, and S. Martín, "Comprehensive Review of Vision-Based Fall Detection Systems," *Sensors*, vol. 21, no. 3, pp. 947-996, 2021, doi: 10.3390/s21030947.

[11] R. Josyula and S. Ostadabbas, "A Review on Human Pose Estimation," *ArXiv*, vol. abs/2110.06877, pp. 1-24, 2021, doi: 10.48550/arXiv.2110.06877.

[12] J.-L. Chung, L.-Y. Ong, and M. C. Leow, "Comparative Analysis of Skeleton-Based Human Pose Estimation," *Future Internet*, vol. 14, no. 12, pp. 380-198,.2022, doi: 10.3390/fi14120380.

[13] L. Mourot, L. Hoyet, F. L. Clerc, F. Schnitzler, and P. Hellier, "A Survey on Deep Learning for Skeleton-Based Human Animation," *Computer Graphics Forum*, vol. 41, no. 1, pp. 122-157, 2021, doi: 10.1111/cgf.14426.

[14] W. W. Y. Ng, M. Zhang, and T. Wang, "Multi-Localized Sensitive Autoencoder-Attention-LSTM For Skeleton-based Action Recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 1678-1690, 2022, doi: 10.1109/TMM.2021.3070127.

[15] S. A. Khowaja and S.-L. Lee, "Skeleton-based human action recognition with sequential convolutional-LSTM networks and fusion strategies," *Journal of Ambient Intelligence Humanized Computing*, vol. 13, no. 8, pp. 3729-3746, 2022, doi: 10.1007/s12652-022-03848-3 .

[16] L. Lu, C. Zhang, K. Cao, T. Deng, and Q. Yang, "A Multichannel CNN-GRU Model for Human Activity Recognition," *IEEE Access*, vol. 10, pp. 66797-66810, 2022, doi: 10.1109/ACCESS.2022.3185112.

[17] G. Weiss, Y. Goldberg, and E. Yahav, "On the Practical Computational Power of Finite Precision RNNs for Language Recognition," *ArXiv*, vol. abs/1805.04908, pp. 1-9, 2018, doi: 10.48550/arXiv.1805.04908.

[18] X. Guo and J. Choi, "Human Motion Prediction via Learning Local Structure Representations and Temporal Dependencies," *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, USA, January 27–February 1, 2019, AAAI Press Publishing, vol. 33, no. 1, pp: 2580-2587, 2019, doi: 10.1609/aaai.v33i01.33012580.

[19] H. Wang, E. S. L. Ho, H. P. H. Shum, and Z. Zhu, "Spatio-Temporal Manifold Learning for Human Motions via Long-Horizon Modeling," *IEEE Transactions on Visualization Computer Graphics*, vol. 27, no. 1, pp. 216-227, 2019, doi: 10.1109/TVCG.2019.2936810.

[20] Y. Li et al., "Efficient convolutional hierarchical autoencoder for human motion prediction," *The Visual Computer*, vol. 35, pp. 1143-1156, 2019, doi: 10.1007/s00371-019-01692-9.

[21] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee, "Convolutional Sequence to Sequence Model for Human Dynamics," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, June 18-22, 2018, pp. 5226-5234, 2018, doi: 10.1109/CVPR.2018.00548.

[22] C. Zang, M. Pei, and Y. Kong, "Few-shot Human Motion Prediction via Learning Novel Motion Dynamics," *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, Yokohama, Japan, July 11-17, 2020, pp. 846-852, 2020, doi: 10.24963/ijcai.2020/118.

[23] M. Al-Faris, J. Chiverton, D. L. Ndzi, and A. I. Ahmed, "A Review on Computer Vision-Based Methods for Human Action Recognition," *Journal of Imaging*, vol. 6, no. 6, pp. 46-77, 2020, doi:10.3390/jimaging6060046.

[24] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-Based Action Recognition with Shift Graph Convolutional Network," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 180-189, doi: 10.1109/cvpr42600.2020.00026.

[25] L. Feng, Y. Zhao, W. Zhao, and J. Tang, "A comparative review of graph convolutional networks for human skeleton-based action recognition," *Artificial Intelligence Review*, vol. 55, pp. 4275-4305, 2021, doi: 10.1007/s10462-021-10107-y.

[26] L. Huang, Y. Huang, W. Ouyang, and L. Wang, "Part-Level Graph Convolutional Network for Skeleton-Based Action Recognition," *Proceedings of the AAAI conference on artificial intelligence*, New York, USA, February 7–12, 2020, AAAI Press Publishing, vol. 34, no. 7, pp. 11045-11052, 2020, doi: 10.1609/AAAI.V34I07.6759.

[27] U. Bhattacharya, T. Mittal, R. Chandra, T. Randhavane, A. Bera, and D. Manocha, "STEP: Spatial Temporal Graph Convolutional Networks for Emotion Perception from Gaits," *Proceedings of the AAAI conference on artificial intelligence*, New York, USA, February 7–12, 2020, AAAI Press Publishing, vol. 34, no. 2, pp. 1342-1350, 2020, doi: 10.1609/aaai.v34i02.5490.

[28] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," *Proceedings of the AAAI conference on artificial intelligence*, New Orleans, Lousiana, USA, February 2-7, 2018, AAAI Press Publishing, vol. 32, no. 1, pp. 1-9, 2018, doi: 10.1609/aaai.v32i1.12328.

[29] W. Peng, X. Hong, H. Chen, and G. Zhao, "Learning Graph Convolutional Network for Skeleton-based Human Action Recognition by Neural Searching," *Proceedings of the AAAI conference on artificial intelligence*, New York, USA, February 7–12, 2020, AAAI Press Publishing, vol. 34, no. 3, pp. 2669-2676, 2020, doi: 10.1609/AAAI.V34I03.5652.

[30] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-Based Action Recognition With Multi-Stream Adaptive Graph Convolutional Networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 9532-9545, 2020, doi: 10.1109/TIP.2020.3028207.

[31] D. Zhang, H. Wang, C. Weng, and X. Shi, "Video Human Action Recognition with Channel Attention on ST-GCN," *Journal of Physics: Conference Series*, IOP Publishing, vol. 2010, no. 1, pp. 012131-012136, 2021, doi: 10.1088/1742-6596/2010/1/012131.

[32] J. Cai, N. Jiang, X. Han, K. Jia, and J. Lu, "JOLO-GCN: Mining Joint-Centered Light-Weight Information for Skeleton-Based Action Recognition," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 2735-2744, doi: 10.1109/WACV48630.2021.00278.

[33] M. Taufeeque, S. Koita, N. Spicher, and T. M. Deserno, "Multi-camera, multi-person, and real-time fall detection using long short term memory," *Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications. SPIE,* vol. 11601, pp. 35-42., 2021, doi: 10.1117/12.2580700.

[34] M. Meratwal, N. Spicher, and T. M. Deserno, "Multi-camera and multi-person indoor activity recognition for continuous health monitoring using long short term memory," *Medical imaging 2022: imaging informatics for healthcare, research, and applications. SPIE*, vol. 12307, pp. 64-71, 2022, doi: 10.1117/12.2612642.

[35] T. Xu, J. Chen, Z. Li, and Y. Cai, "Fall Detection Based on Person Detection and Multi-target Tracking," *11th International Conference on Information Technology in Medicine and Education (ITME). IEEE*, pp. 60-65, 2021, doi: 10.1109/ITME53901.2021.00023.

[36] S. Maldonado-Bascón, C. Iglesias-Iglesias, P. Martín-Martín, and S. Lafuente-Arroyo, "Fallen People Detection Capabilities Using Assistive Robot," *Electronics*, vol. 8, no. 9, pp. 915-934, 2019, doi: 10.3390/ELECTRONICS8090915.

[37] Y. Zhang, J. Yu, Y. Chen, W. Yang, W. Zhang, and Y. He, "Real-time strawberry detection using deep neural networks on embedded system (rtsd-net): An edge AI application," *Comput. Electron. Agric*, vol. 192, pp. 106586-106604, 2022, doi: 10.1016/j.compag.2021.106586.

[38] A. Singh et al., "Artificial intelligence in edge devices," *Advances in Computers*, vol. 127, pp. 437-484, 2022, doi: 10.1016/bs.adcom.2022.02.013.

[39] R. Poojary and A. Pai, "Comparative Study of Model Optimization Techniques in Fine-Tuned CNN Models," *International Conference on Electrical Computing Technologies Applications*, pp. 1-4, 2019, doi: 10.1109/ICECTA48151.2019.8959681.

[40] R. Mishra, H. P. Gupta, and T. Dutta, "A Survey on Deep Neural Network Compression: Challenges, Overview, and Solutions," *ArXiv*, vol. abs/2010.03954, pp. 1-14, 2020, doi: 10.48550/arXiv.2010.03954.

[41] B. L. Deng, G. Li, S. Han, L. Shi, and Y. Xie, "Model Compression and Hardware Acceleration for Neural Networks: A Comprehensive Survey," *Proceedings of the IEEE*, vol. 108, no. 4, pp. 485-532, 2020, doi: 10.1109/JPROC.2020.2976475.

[42] T. Choudhary, V. K. Mishra, A. Goswami, and S. Jagannathan, "A comprehensive survey on model compression and acceleration," *Artificial Intelligence Review*, vol. 53, pp. 5113-5155, 2020, doi: 10.1007/s10462-020-09816-7.

[43] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," Proceedings of the *IEEE/CVF Conference on Computer Vision Pattern Recognition*, Salt Lake City, USA, June 18-22, 2018, pp. 4510-4520, 2018, doi: 10.1109/CVPR.2018.00474.

[44] J. Han and Y. Yang, "L-Net: lightweight and fast object detector-based ShuffleNetV2," *Journal of Real-Time Image Processing*, vol. 18, no. 6, pp. 2527-2538, 2021, doi: 10.1007/s11554-021-01145-4.

[45] J.-H. Kim, S. Chang, and N. Kwak, "PQK: Model Compression via Pruning, Quantization, and Knowledge Distillation," *ArXiv*, vol. abs/2106.14681, pp. 1-5, 2021, doi: 10.48550/arXiv.2106.14681.

[46] M. Zhu and S. Gupta, "To prune, or not to prune: exploring the efficacy of pruning for model compression," *ArXiv*, vol. abs/1710.01878, pp. 1-11, 2017, doi: 10.48550/arXiv.1710.01878.

[47] A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," *ArXiv*, vol. abs/1802.05668, pp. 1-21, 2018,

doi: 10.48550/arXiv.1802.05668.

[48] L. Chen, Y. Chen, J. Xi, and X. Le, "Knowledge from the original network: restore a better pruned network with knowledge distillation," *Complex Intelligent Systems*, vol. 8, pp. 709-718, 2021, doi: 10.1007/s40747-020-00248-y.

[49] Y.-W. Hong, J.-S. Leu, and M. Faisal, "Analysis of Model Compression Using Knowledge Distillation," *IEEE Access*, vol. 10, pp. 85095-85105, 2022, doi: 10.1109/access.2022.3197608.

[50] V. Viswanatha, K. ChandanaR, and C. RamachandraA., "Real Time Object Detection System with YOLO and CNN Models: A Review," *ArXiv*, vol. abs/2208.00773, pp. 1-8, 2022, doi: 10.48550/arXiv.2208.00773.

[51] J. Zheng, H. Wu, H. Zhang, Z. Wang, and W. Xu, "Insulator-Defect Detection Algorithm Based on Improved YOLOv7," *Sensors*, vol. 22, no. 22, pp. 8801-8823, 2022, doi: 10.3390/s22228801.

[52] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 43, pp. 172-186, 2018, doi: 10.1109/TPAMI.2019.2929257.

[53] E. Alam, A. Sufian, P. Dutta, and M. Leo, "Vision-based Human Fall Detection Systems using Deep Learning: A Review," *Computers in biology medicine*, vol. 146, pp. 105626-105664, 2022, doi: 10.1016/j.compbiomed.2022.105626.

[54] E. Auvinet, C. Rougier, J.Meunier, A. St-Arnaud, and J. Rousseau, "Multiple cameras fall dataset," *D.-U. d. Montréal*, Ed., ed, 2010.

[55] I. Charfi, J. Mitéran, J. Dubois, M. Atri, and R. Tourki, "Optimized spatio-temporal descriptors for real-time fall detection: comparison of support vector machine and Adaboost-based classification," *Journal of Electronic Imaging*, vol. 22, no. 4, pp. 041106-041123, 2013, doi: 10.1117/1.JEI.22.4.041106.

[56] K. Adhikari, A. Bouchachia, and H. Nait-Charif, "Activity recognition for indoor fall detection using convolutional neural network," *15th IAPR International Conference on Machine Vision Applications(MVA)*, pp. 81-84, 2017, doi: 10.23919/MVA.2017.7986795.

[57] B. Kwolek and M. Kepski, "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Computer methods programs in biomedicine*, vol. 117, no. 3, pp. 489-501, 2014, doi: 10.1016/j.cmpb.2014.09.005.

[58] A. Shimanoe and M. Amemiya, "Identifying factors of public acceptance for usage of CCTV image," *Journal of the City Planning Institute of Japan*, vol.54, no. 3, pp. 750-757, 2019, doi: 10.11361/journalcpij.54.750.

[59] A. Zatserkovnyy and E. Nurminski, "Identification of Location and Camera Parameters for Public Live Streaming Web Cameras," *Mathematics*, vol. 10, no. 9, pp. 3601-3620, 2022, doi: 10.3390/math10193601.

[60] C.-B. Lin, Z. Dong, W.-K. Kuan, and Y.-F. J. A. S. Huang, "A Framework for Fall Detection Based on OpenPose Skeleton and LSTM/GRU Models," Applied Sciences, vol. 11, no. 1, pp. 329-348, 2020, doi: 10.3390/app11010329.

[61] L. Zhao and L. Wang, "A new lightweight network based on MobileNetV3," *KSII Transactions on Internet Information Systems*, vol. 16, no. 1, pp. 1-15, 2022, doi: 10.3837/tiis.2022.01.001.

[62] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 42, pp. 2011-2023, 2018, doi: 10.1109/TPAMI.2019.2913372.

[63] A. G. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *ArXiv*, vol. abs/1704.04861, pp. 1-9, 2017, doi: 10.48550/arXiv.1704.04861.

[64] E. Kurniawan et al., "Deep neural network-based physical distancing monitoring system with tensorRT optimization," *International Journal of Advances in Intelligent Informatics*, vol. 8, no. 2, pp. 1-16, 2022, doi: DOI:10.26555/ijain.v8i2.824.

[65] L. Liu, E. B. Blancaflor, and M. B. Abisado, "A Lightweight Multi-Person Pose Estimation Scheme Based on Jetson Nano," *Applied Computer Science*, vol. 19, no. 1, pp. 1-14, 2023, doi: 10.35784/acs-2023-01.

[66] M. S. Pavithra, K. Saruladha, and K. Sathyabama, "GRU Based Deep Learning Model for Prognosis Prediction of Disease Progression," in *3rd International Conference on C*omputing Methodologies Communication, vol. 2019, pp. 840-844, 2019, doi: 10.1109/ICCMC.2019.8819830.

[67] A. Carlier, P. Peyramaure, K. Favre, and M. Pressigout, "Fall Detector Adapted to Nursing Home Needs through an Optical-Flow based CNN," *42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society*, vol. 2020, pp. 5741-5744, 2020, doi: 10.1109/EMBC44109.2020.9175844.

[68] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View Adaptive Neural Networks for High Performance Skeleton-Based Human Action Recognition," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 41, no. 8, pp. 1963-1978, 2019, doi: 10.1109/TPAMI.2019.2896631.

[69] I. Kareem, S. F. Ali, and A. Sheharyar, "Using Skeleton based Optimized Residual Neural Network Architecture of Deep Learning for Human Fall Detection," in *IEEE 23rd International Multitopic Conference*, vol. 2020, pp. 1-5, 2020, doi: 10.1109/INMIC50486.2020.9318061.

[70] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "Skeleton-based human activity recognition using ConvLSTM and guided feature learning," *Soft Computing*, vol. 26, no. 2, pp. 877-890, 2021, doi: 10.1007/s00500-021-06238-7.

[71] Y. Zheng, D. Zhang, L. Yang, and Z. Zhou, "Fall detection and recognition based on GCN and 2D Pose," *6th International Conference on Systems Informatics*, IEEE, vol. 2019, pp. 558-562, 2019, doi: 10.1109/ICSAI48974.2019.9010197.

[72] J. Lee and S. J. Kang, "Skeleton action recognition using Two-Stream Adaptive Graph Convolutional Networks," *36th International Technical Conference on Circuits/Systems, Computers Communications (ITC-CSCC)*. IEEE, vol. 2021, pp. 1-3, 2021, doi: 10.23919/CCC55666.2022.9901587.

[73] B. U. Maheswari, R. Sonia, M. P. Rajakumar, and J. Ramya, "Novel Machine Learning for Human Actions Classification Using Histogram of Oriented Gradients and Sparse Representation," Information Technology and Control, vol. 50, no. 4, pp. 686-705, 2021, doi: 10.5755/j01.itc.50.4.27845.

[74] S. Kiran *et al.*, "Multi-Layered Deep Learning Features Fusion for Human Action Recognition," *Computers Materials & Continua,* vol. 69, no. 3, pp. 1-15, 2021. doi: 10.32604/cmc.2021.017800.

Lei Liu

Lie Liu was born in Anhui, China, in 1987. He received his B.S. degree from the Anhui University, in 2010, the M.S. degree from Hefei University of Technology, in 2013, and PhD degree from the National University, Philippines, in 2023. He is currently a lecturer at the School of Computer Science, Huaian Normal University. His main research interests include computer vision and machine learning.

Yeguo Sun

Yeguo Sun was born in Anhui, China, in 1979. He received his B.S. degree from the Department of Mathematics and Computational Science, Fuyang Normal University, Fuyang, China, 2001, M.S. degree from the Department of Mathematic-s, Shanghai Normal University, Shanghai, China, 2007, and PhD degree from the Department of Control Science and Engineering, Beijing University of Aeronautics and Astronautics, Beijing, Chi-na, 2011. He is now a Professor at the School of Finance and Mathematics in Huainan Normal University. He has published several papers and some of them were indexed by SCI and EI. His current research interests include network control systems, neural networks, finite-time control.

Xianlei Ge

Xianlei Ge received his B.S. and M.S. degrees in information communication engineering from Chongqing University of Posts and Communications, Chongqing, China, in 2012 and 2015, respectively. From 2015 to 2016, he worked as a software engineer at Huawei Technologies Co., Ltd. He is a lecturer and researcher at Huainan Normal University since 2016. He is currently pursuing his Ph.D. degree in computer science at the College of Computing and Information Technologies, National University in Manila, Philippines. His research interests include image processing, machine learning and natural language processing.

Spatial-Aware Multi-Level Parsing Network for Human-Object Interaction

Zhan Su[1], Ruiyun Yu[1]*, Shihao Zou[2], Bingyang Guo[1], Li Cheng[2]

[1] Software College, Northeastern University, Shenyang (China)
[2] University of Alberta, Edmonton (Canada)

* Corresponding author: yury@mail.neu.edu.cn

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

Human-Object Interaction (HOI) detection focuses on human-centered visual relationship detection, which is a challenging task due to the complexity and diversity of image content. Unlike most recent HOI detection works that only rely on paired instance-level information in the union range, our proposed Spatial-aware Multi-level Parsing Network (SMPNet) uses a multi-level information detection strategy, including instance-level visual features of detected human-object pair, part-level related features of the human body, and scene-level features extracted by the graph neural network. After fusing the three levels of features, the HOI relationship is predicted. We validate our method on two public datasets, V-COCO and HICO-DET. Compared with prior works, our proposed method achieves the state-of-the-art results on both datasets in terms of mAP_{role}, which demonstrates the effectiveness of our proposed multi-level information detection strategy.

Keywords

I. Introduction

IMAGES are the main form of information obtained by humans. In recent years, basic vision tasks, such as target detection, action recognition and image segmentation, have developed rapidly with the application of deep learning. Research on higher-level image semantics of individual instances, such as human action recognition and pose estimation, has also made significant progress. Human-Object Interaction (HOI) detection, an intersecting area of object detection [1], behavior recognition [2], and visual relationship detection [3], utilizes images as input to detect and locate human-object pairs and predict their interaction categories. Formally, Visual Relationship Detection (VRD) uses ⟨*objectA, predicate, objectB*⟩ to define relational expressions, which involves a combination of interactions of multiple target objects, such as human-human, human-object, object-object, etc. HOI detection only focuses on the interaction between humans and objects, and the predicates are mainly concentrated in the category of verbs, which have significant reference value for the development of behavior recognition. The diversity of the interaction between humans and objects mainly relies on the object category, the human's pose, and the human's relative position with the object. In some cases, even if the object, pose, and relative position are the same, the behavior could be different. For example, putting things back and picking things up are two different behaviors. These situations make the HOI a challenging task [4].

For the HOI detection task on static images, Gupta and Malik [5] first solved this problem. Before their work, most researchers only recognize human actions and bounding boxes in a single or multiple frames. Based on object detection, they associate various semantic instances in the scene, detect the performed actions and recognize the interactions between object instances and humans, which provides a detailed understanding of current activities. Georgia Gkioxari et al. [6] propose a human-centric architecture strategy for detecting all interactions between objects and humans, considering the number of objects in an image is unknown. The mainstream methods in HOI detection tasks [7]–[10] adopt a similar human-centered network model structure. Gao et al. [7] propose the human-centered attention module iCAN to emphasize the contextual information areas in the image related to interaction. The main idea of this module is to utilize the softmax function to transform the fused instance-level appearance features and convolutional features to an attention map, which produces high-level features. Wang et al. [8] improves the iCAN module by embedding the context-aware appearance and attention modules in the "human stream" and "object stream" to extract the appearance and context information in the image. Bansal et al. [9] proposes the spatial priming model to strengthen the relative spatial position between humans and objects. Liao et al. [10] propose a single-stage parallel point detection and matching model, where the point detection branch estimates human points, interaction points and object points, and the point matching branch takes the human and object points with corresponding interaction points as a matching pair. This

method screens out many candidates interaction points to solve the large model scale, increasing detection speed. Although these methods have achieved significant improvement in HOI detection, they do not consider the spatial configuration relationship between all objects and humans in the scene at once and do not make full use of the pose and part-level information of human instances.

The HOI of a single interaction pair is related to its spatial positions, the global interaction relationship with other instances, the pose, and specific parts of a human body. In order to obtain more information, we propose a Spatial-aware Multi-level Parsing Network (SMPNet) for HOI detection. Specifically, an image is an input to the backbone of Faster RCNN [11] and CPN [12] to obtain the feature maps and human pose and then fed into the multi-branch deep network to perform relational inference based on the multi-level features of each human-object pair. Fig. 1 shows an example of our relation reasoning.
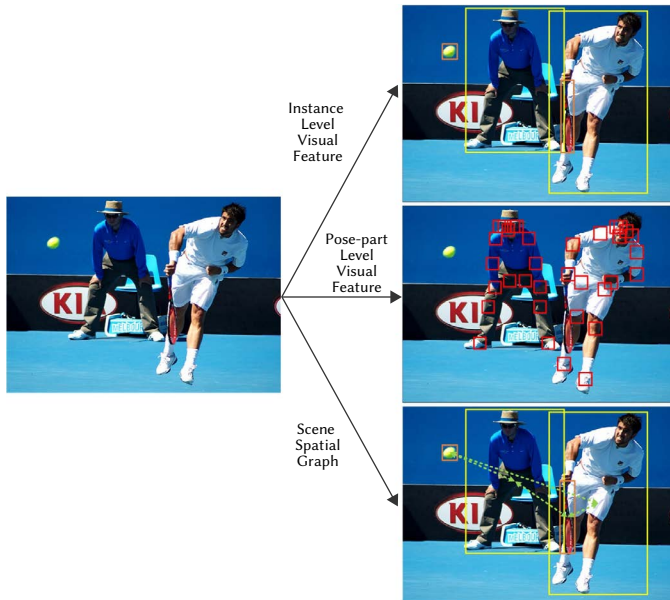
Fig. 1. SMPNet uses three levels of information, including i) the visual features of humans (yellow boxes), objects (orange boxes), and the overall context of an image, ii) the visual features of various parts of the human body (red boxes), iii) the information propagated on the scene spatial graph (green lines).

SMPNet consists of an instance-level and a part-level visual branch to process the instance-level and part-level appearance features, respectively, which are further refined by the spatial attention scores and features from the spatial configuration map. The pose key-points of the target human are obtained through the pose estimator. Then, we embed the encoded pose features into the spatial configuration features within the human bounding box. At the same time, we use the key-points as the center point to obtain the part-level features. Then, the attention weight is extracted to achieve refined instance-level and part-level features. Although some works [7]–[9] have used these spatial configurations directly as classification features, they do not combine pose information with spatial information, and therefore neglect it as a clue to infer part-level human features. Our approach provides attention mechanisms for refining visual features from multiple levels.

In the spatial graph branch, to effectively utilize the spatial configuration information of all instances in the entire scene, we model the scene in the image as a graph, the nodes of which are all the detected humans and objects, and integrate the spatial configuration information to define the propagation on the edges. In this case, each human node can receive lots of messages indicating the existence of other object nodes. If we assume the messages sent from a single object node to all human nodes are the same, then the only variable

is the propagation weights that control the information propagation. On this basis, each human node can receive information about the objects' existence and their relative position. After information propagation, node features can effectively integrate each instance's visual and spatial location information in the image and provide additional scene-level clues for interactive detection. Finally, we can predict the HOI categories for each sample based on fused features.

Using the V-COCO [5] dataset and the HICO-DET [13] dataset, we conduct extensive experiments. The observed results reveal that SMPNet surpasses the state-of-the-arts. In general, the following are the major contributions of this research:

- To integrate and extract the correlation between interactive instances in the global environment, we present a special graph neural network design.
- We use the spatial configuration map containing the pose information to obtain the visual appearance features' attention weight and refine the potential multi-level features.
- We use the encoded spatial features combined with the visual appearance features to obtain the message so that the information propagation in the graph neural network process is adjusted according to the interaction pair situation and the sender situation.

The rest of this article is organized as follows. Section II introduces the related work. Section III explains the SMPNet model comprehensively. Extensive experiments are executed in Section IV to demonstrate the efficacy of our method. Lastly, Section V concludes the article.

II. Related Work

A. Attention Mechanism

The attention mechanism mainly utilizes human vision to quickly scan the global image and then obtain the target area in the image that needs attention. This idea is to imitate the special brain signal processing mechanism of human vision and appropriately invest more attention in the target area to obtain more detailed information and suppress other useless information. Attention mechanism (AM) mainly builds an attention matrix to make the deep neural network pay attention to the key features in the image during the training process to avoid the impact of non-key features. The attention mechanism is first applied in machine vision, and its main function is to make the areas that need to be focused on in the data to get more attention [14]–[19], Bahdanauu et al. [18] apply the attention mechanism in the machine translation task. Their work is further verified that the attention mechanism can effectively reflect the relationship between features, promoting deep neural networks combined with attention mechanism research. Subsequently, VaSWanl et al. [19] introduce the attention mechanism into the sentence modeling task and use a two-dimensional matrix to represent sentence information, thereby obtaining a feature representation of richer semantic information. Later, researchers introduce the attention mechanism into image processing, such as Gao et al. [7]. Our task focuses on part-level and instance-level attention for HOIs detection.

B. Object Detection

Traditional target detection algorithms can be divided into target instance detection and traditional target class detection. The former considers that the objects in the image are irrelevant except for the specific target of interest. The detection target instance usually uses the template and image stability features to obtain correspondence between the objects in the scene. The latter is based on the selected features and classifiers using HOG [20] features, support vector machine [21] and AdaBoost [22] algorithm framework and other methods to detect a limited number of classes.

Alex et al. propose the Alexnet convolutional neural network model and improve it significantly. Compared with traditional algorithms that manually extract features, deep neural networks have considerably improved in nature. Since then, deep neural networks' powerful autonomous learning and expression capabilities have replaced traditional feature extraction methods. Target detection methods based on deep learning include two-stage target detection algorithms and one-stage target detection algorithms. RCNN [23], Fast RCNN [24], Faster RCNN [11], etc., are common two-stage target detection algorithms, and YOLO [25], SSD [26], etc., are common one-stage target detection algorithms. The two-stage target detection algorithm uses a convolutional neural network to classify the generated candidate frame samples. The one-stage target detection algorithm is different from the two-stage target detection algorithm. It does not need to generate a candidate frame but directly converts the problem of target frame positioning into a regression problem. Additionally, we utilize the Faster RCNN with ResNet-50-FPN as a region proposal network and extend it to interaction proposals that predict if a human-object pair is interacting.

C. Visual Relationship Detection Technology

Image understanding can identify the relationship between objects in an image and form a comprehensive language description. It is one of the widest applications in image processing. In general, an image usually contains multiple interacting objects. Recognizing a single object is not enough to better understand these images. The relationship between objects also contains very important information. And sometimes, this relationship determines the semantic information represented by this image. This has led to the diverse relationships in the image becoming the focus and difficulty of image understanding.

Visual relationship detection technology is mainly divided into visual relationship detection based on scene graphs and visual relationship detection based on visual features. The former utilizes scene graphs to understand images' high dimensional semantic meaning as a problem of obtaining a directed graph structure. Johnson et al. [27] first propose the concept of Scene Graphs, which can more accurately understand the semantic information of images. Jianwei Yang et al. propose the Graph-R CNN framework [28], which utilizes graph convolution based on the directed graph structure to identify the relationship between objects. The relation proposal network (RePN) model they proposed effectively solves the problem that the connection between two objects increases with the number of objects squared. In addition, they also propose an attention graph convolutional network (aGCN), which can efficiently obtain objects' interrelationships between objects instances. The latter's representative is the visual transformation embedding network (VTransE) [29] proposed by Hanwang Zhang et al. for visual relationship detection. The characteristic of the network is to put the target in the designed low-dimensional relational space and utilize simple vector transformation in this space to express the relation between objects. For example, the subject&predicate are approximately equal to the object. At he same time, they utilize a novel feature extraction layer, which enables the transfer of target relationship knowledge in the form of full convolution and trains in a simple forward or backward. In our research, our focus is on HOI detection, which is a human-centric problem, to detect action interactions between humans and objects.

D. HOI Detection

Based on the human-centric network model and the human-object region convolutional neural network (HO-RCNN), Chen Gao et al. propose an end-to-end trainable attention network iCAN [7]. First, the network utilizes humans as the center of interaction to obtain the position probability map of the object interacting with the human. Through the position probability map, they clarify the relationship pair.

Based on the iCAN model, many variants using mixed approaches [8], [30]–[32] have been proposed. Wang et al. [8] proposed a contextual attention model, which can adaptively learn the context information of instances, allowing the network to focus on important semantic areas. Xu et al. [30] constructed a priori-knowledge graph based on the annotation of the HOI dataset and the external visual relationship dataset and modeled the semantic relationship between verbs and objects to alleviate the long-tail distribution problem. Wang et al. [31] started from the region proposal module, proposed an HO-RPN network suitable for HOI detection, and introduced an external word embedding in the object classification to achieve Zero-Shot Learning. Bansal et al. [32] designed the functional generalization module, which uses the word embedding of human and object as a feature to endow the model with zero-shot learning capabilities. Song Gao et al. [33] improve the accuracy of the existing model by optimizing the loss function and training details to improve the performance of the HOI detection task. However, those works only take instance-level features of humans and objects but do not utilize the features in each part of the human, and other scene instances with spatial information, which provides more detailed information for the HOI task. We propose a model that captures multi-level information from input feature maps.

III. Spatial-Aware Multi-Level Parsing Network

This section introduces our proposed SMPNet for detecting human-object interaction. As shown in Fig. 2, we integrate the features and values of three individual branches to perform interaction prediction: the first is utilized to analyze human part-level visual features, the second is applied to analyze instance-level visual features, and the third is used to extract interactive features between all instances in the scene as scene-level features. The spatial configuration is an important reference feature that indicates whether an object interacts with a human. For example, when the object and the human proposals are close in space, or even the object proposal is close to the spatial position of any pose key-point, they usually have direct visual interaction. In contrast, there is usually no direct visual interaction when two proposals are spatially far apart. According to this property, introducing a spatial configuration map to generate weights can better refine the weights of instance-level and part-level features. Then, the spatial features between the interaction pairs are coded together with the appearance features and used as the input to the spatial graph branch, and the interaction features are obtained through message passing of the graph neural network. This branch considers multiple groups of interactions involving multiple humans and multiple objects in the scene. These interaction features can provide additional reference features for detecting human-object pairs. For example, the content in the image is a concert, and the information that other humans are playing some musical instruments in the image can provide reference features for the currently detected human-object pair playing the piano.

We extract the appearance features of instances and context (Sec. A) based on the work of Chen et al. [7]. Following that, we explain the fundamental method that incorporates the spatial configuration's attention mechanism, including the pose information into this branch network. We use human pose cues to improve the local semantics on the spatial configuration map and extract attention scores (Sec. B). Then, we utilize these scores to enhance the visual appearance features of part-level. Next, we obtain relevant features within instances based on a bipartite scene graph (Sec. C). At last, we describe the fusion of instance-level visual features, human part-level visual features, and scene-level instance-related features (Sec. D).
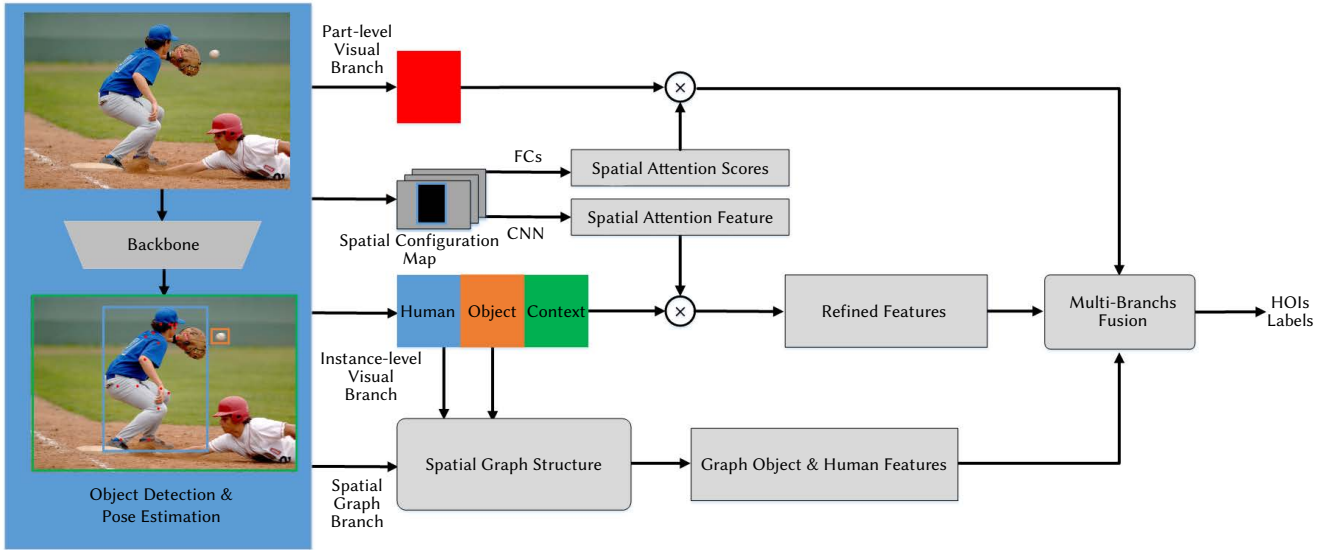
Fig. 2. Overview of our framework: For an interaction pair (a human and an object), the "backbone module" aims to prepare convolutional feature maps and the ROIs for three parallel branch networks. Rounded rectangles are operations, and ⊕ is element-wise multiplication.

A. Instance-Level Visual Branch

We construct a branch network based on the method in [7] to extract instance-level visual appearance features, including human area, object area, and scene context. Detailed information is shown in Fig. 3.
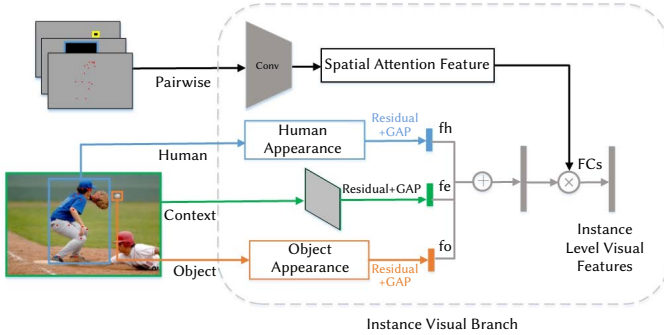
Fig. 3. Structure of the instance level visual branch. The branch contains human, object, context, and pairwise spatial streams. Here ⊕ is the concatenation process, ⊗ represents element-wise multiplication, GAP is global average pooling, Residual denotes residual block [34], and FCs denotes two fully connected layers.

This branch focuses on extracting the visual appearance features of human-object pairs. Referring to the multiple variants of the iCAN model using mixed approaches [8],[30]–[32], we designed the instance-level visual branch that includes four proven effective feature streams: object, human, context, and pairwise streams. Compared to [7], we utilize RoIAlign rather than ROI pooling and adjust the dimensions of each part. RoIAlign is a method proposed in Mask-RCNN [35] to aggregate and output specified size features in regions of different sizes on the feature map. It uses the bilinear interpolation method to obtain the feature values on the pixels whose coordinates are floating-point numbers, thus transforming the whole process of feature aggregation into a continuous operation. We use RoIAlign on the human and object regions to extract features following object detection. This operation is followed by a residual block (Res) [34] and global average pooling (GAP) to extract visual feature vectors of objects, humans and context.

In contrast to the late fusion approach in [7], we apply the early fusion approach to process the features of the human, object, and context features as f_h, f_o, f_e, which is to concatenate all the features and project it to obtain the instance-level visual appearance feature as equation (1):

$$f_{ivis} = W_{ivis}(f_h \oplus f_o \oplus f_e) \tag{1}$$

where ⊕ denotes concatenation operation, W_{ivis} is the projection matrix which is realized through a fully connected layer, f_{ivis} is a feature vector of size D.

To focus on learning the spatial interaction mode between humans and objects, the output of the pairwise stream is the attention feature. This attention feature is used to refine the visual features of humans, objects, and the context obtained in the other three streams, as shown in Fig. 3.

We use the two binary masks of humans and objects proposal in the image as clues to capture the instance-level spatial configuration, similar to [7], [8], [36]. In detail, according to the given human proposal x_h and the object proposal x_o, we generate two binary images. These binary maps have zeros everywhere except for the region defined by the human and object proposal x_h and x_o of each map, respectively. At the same time, to match the pose-parts' visual features, we encode the pose key-points extracted in the backbone stage into a map according to the skeleton configuration of the coco dataset following the work of Yong-Lu Li [37]. In this map, the key-points are connected by lines with diverse gray values between 0.1 and 0.9, indicating the corresponding parts of the pose, and the values of the remaining areas are all 0. After that, the two kinds of maps are rescaled to the size of $M \times M$ and connected to generate a 3-channel binary scene spatial configuration map M_{hop}.

Referring to the work of [7], [13], we utilize two convolutional layers to parse the scene spatial configuration map. The following equation (2) is the GAP and full connection operation:

$$a_{hop} = W_{hop}(GAP(Conv(M_{hop}))) \tag{2}$$

where W_{hop} denotes a fully connected layer and a_{hop} is the attention feature vector of the same size as the visual feature vector f_{ivis} obtained in the previous branch.

On this basis, we utilize a_{hop} to refine the visual feature vector f_{ivis} through the vector multiplication operation to obtain the output f_a of this branch. The formula (3) is as follows:

$$f_a = f_{ivis} \otimes a_{hop} \tag{3}$$
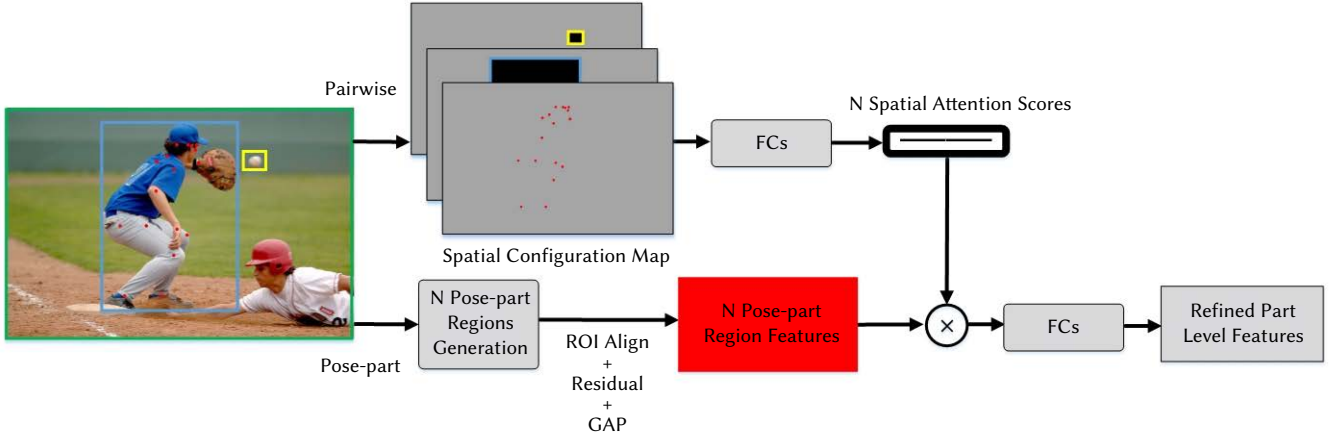
Fig. 4. Structure of the part-level visual branch. The overall module includes pose-part and pairwise spatial streams. Here $\oplus$ denotes concatenation operation, $\otimes$ is element-wise multiplication, Residual denotes block [34], GAP is global average pooling, and FC denotes fully connected layer.

where $\otimes$ is the element-wise multiplication, and f_a is the refined feature vector of size D.

B. Part-Level Visual Branch

As shown in Fig. 4, this branch focuses on extracting the visual features of the part-level. First, we extract the part area of each key-point representing N human body parts. From the backbone stage, we obtain all the pose key-points $k(p)$ of the human h. Every pose point serves as the center for the generation of an area $R_{pk} = \{h_{pk}, w_{pk}, x_{pk}, y_{pk}\}$. Furthermore, during the process of generating the area, it is stipulated that it will not exceed the range of the image. The computation procedure is as equation (4):

$$h_{pk} = w_{pk} = \lfloor \delta \sqrt{h_{human} * w_{human}} \rfloor \qquad (4)$$

where w_{human} and h_{human} indicate the size parameters of the human bounding box, the notation [] denotes the rounding-up process, and δ indicates the scaling value that is adjusted to 0.1 based on an experimental evaluation.

Then, based on the resulting pose-part areas $R_{parts} = \{R_{p1}, \ldots, R_{pN}\}$, we utilize the RoIAlign algorithm and GAP operation extracts N ROI features $F_{key} = \{f_{p1}, f_{p2}, \ldots, f_{pN}\}$ for each part on shared feature maps with deviation information. We encode deviation information with a two-channel feature map, where the channels denote the x and y offsets of each pixel on the feature map to the center point of the object bounding box. Therefore, we then connect it with the image feature map. We utilize two connected layers to parse the scene spatial configuration map to get spatial attention scores $B_{hp} = \{b_{hpi}\} \in \mathbb{R}^N$. ReLU layer is adopted after the first layer, and a Sigmoid layer is used after the second layer to normalize the final prediction to [0,1]. The scores are utilized for weighting the pose-part area features, and the output feature f_b of this branch is got via the concatenation process and two layers of full connection, as shown in equation (5):

$$f_b = FCs((b_{hp1} \otimes f_{p1}) \oplus \ldots \oplus (b_{hpN} \otimes f_{pN})) \qquad (5)$$

where $\{b_{hpi} \in [0,1]\}_{i=1}^N$, $\oplus$ indicates the concatenation process, $\otimes$ indicates element-wise multiplication, $\{f_{pi}\}_{i=1}^N$ mean the pose-parts features with deviation information and FCs represents two fully connected layers.

C. Spatial Graph Branch

To generate effective features containing visual and spatial information for human-object pairs, in this branch, we propose to use humans and objects as nodes and their relationships as edges to construct a graph structure and utilize graph neural networks for parsing.

We construct a bipartite graph $\mathcal{G} = (\mathcal{H}, \mathcal{O}, \mathcal{E})$ instead of a fully connected graph. This simplification is to avoid unnecessary calculations. One side of the graph is the p_i set H with $c_i = h$, and the other is the p_i set O with $c_i \neq h$, where h indicates the object is human or not. Edge set $\mathcal{E}$ connects all the detected humans and objects, as shown in Fig. 5. Instance-level visual appearance features and spatial information features determine the weight of the edge between the interaction pairs. The parsing module generates the feature message for propagation in the graph neural network according to the sender's visual appearance features and the encoded spatial information features. To construct the graph, we use Faster RCNN [11] as an object detector and apply appropriate filtering to obtain the detection candidate set $\{p_i = (b_i, s_i, c_i)\}_{i=1}^n$. b_i denotes the bounding box with dimension 4, si P r0, 1s represents the bounding box with dimension 4, $s_i \in [0, 1]$ represents the score given by the detector, and $c_i \in C$ represents the object category of the candidate. We utilize RoIAlign [35] to extract the visual features of the candidate set as node features, and the calculated spatial feature vectors are used as edge features.
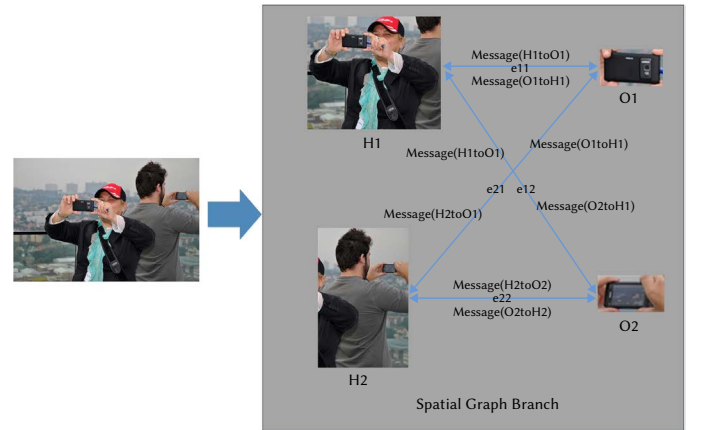
Fig. 5. Spatial graph branch. This branch takes humans and objects as two kinds of nodes to form a bipartite graph structure.

We calculate the human-object information in the image space as the spatial feature, including the center coordinates, width, height, aspect ratio, intersection and area of human and object bounding boxes, and normalize by the image size. d_x is the difference between the center point coordinates on the horizontal axis. d_y is the difference of the center point coordinates on the vertical axis. After normalization by the size of human bounding box, we utilize $[ReLU(d_x), ReLU(-d_x), ReLU(d_y), ReLU(-d_y)]$ to express the orientation of the bounding box

pair. Then, we obtain the spatial information feature vector f_{sp} with dimension 18. Regarding the work of Gupta et al. [38], we connect it with its logarithm to learn higher-order combinations of different terms as equation (6):

$$f'_{sp} = f_{sp} \oplus log(f_{sp} + \theta) \tag{6}$$

where θ is a small constant greater than 0, $\oplus$ denotes concatenation operation, and log represents a logarithmic operation. Then, we utilize a fully connected layer to transform f'_{sp} to the same dimensional space as f_h and f_o.

We adopt the graph neural network and propagate the feature message determined by the sender and spatial information features. Therefore, even if it is the same sender, the feature message will differ depending on the receiver. The message parsing module is defined as equations (7) and (8):

$$M_{oh}(f_{oj}, e_{ij}) = FCs(f_{oj} \oplus e_{ij}) \tag{7}$$

$$M_{ho}(f_{hi}, e_{ij}) = FCs(f_{hi} \oplus e_{ij}) \tag{8}$$

where $\oplus$ represents the concatenation process, and FCs represents two layers of full connection. The weights of the human node as the sender and the object node as the sender are not shared. f_{oj} is the node feature of the j-th object ($j \in \{1, \ldots, |\mathcal{O}|\}$), f_hi the node feature of the i-th human ($i \in \{1, \ldots, |\mathcal{H}|\}$), e_{ij} is f'_{sp} of the pair of the i-th human and the j-th object.

According to the given feature message and edge weights, the process of transferring feature messages in the graph structure is defined as equations (9) and (10):

$$f'_{hi} = f_{hi} + \sum_{j=1}^{\mathcal{O}} \alpha_{ij} M_{oh}(f_{oj}, e_{ij}) \tag{9}$$

$$f'_{oj} = f_{oj} + \sum_{i=1}^{\mathcal{H}} \alpha_{ji} M_{ho}(f_{hi}, e_{ij}) \tag{10}$$

where α_{ij} and α_{ji} denote the weight relationship between h_i and o_j. Different from previous works [39], [40], the weight of edges is defined as the visual similarity. In our proposed method, the weight of edges is determined by instance-level visual features and spatial information features. Therefore, the edge weight value of the pair of the i-th human and the j-th object is calculated as equation (11):

$$\alpha_k = FCs(f_{hi} \oplus f_{oj} \oplus e_{ij}) \tag{11}$$

where $\oplus$ represents concatenation operation, and FCs represents two fully connected layers, $k \in \{1, \ldots, |\mathcal{O} \times \mathcal{H}|\}$. A ReLU layer and Sigmoid function are used following the first and second layers. The edge weight value α_{ij} is obtained by applying softmax according to the number of connected object nodes and α_k. Similarly, α_{ji} is normalized according to the number of connected human nodes and α_k. Finally, we pair the graph features and calculate the output features f_c as equation (12):

$$f_c = FCs(f'_{hi} \oplus f'_{oj}) \tag{12}$$

where $\oplus$ represents the concatenation process, and FCs represents two layers of full connection.

D. Training and Inference

We obtain the triples of ⟨human (h), interaction (i), object (o)⟩ to calculate the HOI score $S^i_{h,o}$ as the final output of our model. We evaluate triples from two aspects: first, use instance-level visual features and context features to quantify whether the human-object has a relationship with the value of the relationship score s_r; second, a human usually executes multiple interactions as a multi-label classification

problem. So we aggregate extra information from scene-level and part-level visual branches, and apply the Sigmoid function to each interaction category to obtain scores $S_{ps} = \{s^i_{ps}\}$ $\left(\forall i \in \mathcal{J} = \{1, \ldots, M\} \right)$, where $\mathcal{J}$ is a collection of interaction categories, and M denotes the total of interaction categories. The two-stage fusion strategy utilizes s_r to inhibit considerable background pairings and enhance the detection precision. They are calculated as equations (13) and (14):

$$s_r = \sigma(FCs(f_a)) \tag{13}$$

$$S_{ps} = \sigma(FCs(f_a \oplus f_b \oplus f_c)) \tag{14}$$

Here, $\oplus$ represents the concatenation process, and σ is a sigmoid function that follows the two layers of full connection. The result features of the three branches are denoted by f_a, f_b, and f_c, respectively.

We utilize the humans and objects with high scores obtained in the object detection stage to form an initial bipartite graph during the training process. Therefore, we combined the two evaluation indicators described above with the detection scores of the human-object pair to obtain the HOI scores $S^i_{h,o}$, as shown in equation (15):

$$S^i_{h,o} = s_r \cdot s^i_{ps} \cdot s_h \cdot s_o. \tag{15}$$

Here, s_h and s_o are the scores of humans and objects obtained in object detection stage.

Because the HOI tags in the different datasets are unbalanced, we utilize the weighted binary cross-entropy loss $\mathcal{L}_{cls} = w_p \cdot y \cdot log(\hat{y}) + w_n \cdot (1-y) \cdot log(1-\hat{y})$, where w_p and w_n denote the weight ratios for positively and negatively samples. Our loss function expression is as equation (16):

$$\mathcal{L} = \sum_{i=1}^{C} \mathcal{L}_{cls}(t^i, s^i_{ps}) + \lambda \mathcal{L}_{cls}(r, s_r) \tag{16}$$

where λ represents the weight used to regulate the impact of the loss term, $T = \{t^i \in [0,1]\}$ denotes the interaction categories label collection, and r represents the interaction relationship label collection. Moreover, t^i denotes the ground truth relationship label of the i-th interaction of the sample, and $r \in \{0,1\}$ represents the presence of the interaction of the pairing.

IV. Experiments

In this part, we describe our experimental result. We begin by explaining the datasets and evaluation metrics along with our implementation details. Then, we perform an extensive quantitative analysis of our proposed model to prove the effectiveness of our approach. Eventually, we utilize ablation experiments to illustrate the influence of these branches in our approach.

A. Datasets and Metric

Datasets. To evaluate the performance of our model, we use two benchmarks for HOI detection: V-COCO [5] and HICO-DET [13]. V-COCO has derived from MS-COCO [41] dataset. It has 10,346 images and 16,199 human instances (2533 images are contained in the training set, 2867 images are for validating, and 4,946 in the test set). The V-COCO dataset contains 26 binary interaction categories. If the object in the image is related to the action, the object is also be annotated. HICO-DET contains 38,118 training images and 9,658 test images with bounding box annotations, 600 HOI categories for 80 object classes (the same as those in the MS COCO data set [41]) and 117 action verbs.

Evaluation metric. For these two data sets, we adopt the evaluation settings in [5]. The results are reported in terms of role mean average precision (*mAProle*). In the research, the purpose is to detect the triples of ⟨*human*, *interaction*, *object*⟩. A detected triplet is deemed as

a true positive if it has the correct action label, and the minimum of human overlap IOU_h and object overlap IOU_o is greater than 0.5. To demonstrate the effectiveness of our proposed method in interactions with different numbers of annotations, we follow previous practices [13], and the report is divided into three different HOI category sets for the HICO-DET dataset: (a) all 600 HOI categories in HICO (Full), (b) 138 HOI categories with less than ten training instances (Rare), and (c) 462 HOI categories with ten or more training instances (Non-Rare).

B. Implementation Details

As previously stated, we use Faster RCNN [11] and ResNet-50-FPN [42] backbone to obtain the bounding box prediction of humans and objects and CPN [12] to estimate human pose. These have been pre-trained using the MS-COCO dataset. The pose structure comprises N = 17 key-points, which correspond to the MS-COCO data set [41]. The ROI feature with the highest resolution is obtained from the feature map in FPN [42]. In the object detection stage, first, we remove the candidate boxes with a score lower than 0.2 and perform a non-maximum suppression (NMS) operation with a parameter of 0.5. After that, we sort the candidate boxes and select the top 15 humans and objects, respectively, to form a bipartite graph and remove the candidate pairs that contain the same human twice. The resolution of the RoIAlign algorithm in the instance level visual branch is $R_h = 7$, through a residual block, and then global average pooling is similar to [7]. After these steps, we obtain three feature vectors of human, object and context with size $R = 256$, size $D = 3R$. In the part level visual branch, the RoIAlign algorithm produces $5 \times 5 \times (R + 2)$ output features for every region and then utilizes a residual block and GAP to downsize it to $1 \times 1 \times (R + 2$ size. To train the model, we use SGD as the optimizer, with a momentum of 0.9 and weight decay of 1e-4. All data sets have an initial learning rate of 4e-2. Our model has trained 36k iterations and 250k iterations on V-COCO and HICO-DET, respectively. Furthermore, for these two data sets, we reduced the learning rate to 4e-3 at iteration 18k and iteration 200k, respectively.

C. Results

Quantitative results on the two test datasets and the performance comparison between our approach and the current approaches are shown in Tables I and II. We set the baseline to include only the four streams in the instance level visual branch. Our final method integrates all the branches and components introduced in section III.

From Table I we can see that we compare our model with the current ten approaches [5]–[8], [40], [43]–[47] on V-COCO dataset. In the existing work, the GPNN method [40] utilizes the graph neural network to learn and detect interactions, reaching a mAP_{role} of 44.0.

The iCAN model [7] integrates three visual feature streams using the attention mechanism in the early fusion approach and provides a mAP_{role} of 45.3. RPNN [44] utilizes a wide range of part-level visual features to detect interactions and realizes a mAP_{role} of 47.53. Zhou et al. [46] introduces a cascade architecture for HOI understanding from coarse to fine and achieves a mAP_{role} of 48.9. PMFNet [47] further divides the part-level visual feature into the same number of pose keypoints and obtains a mAP_{role} of 52.0. Our baseline method achieves 49.3 mAP, and the full method performs the highest effect of 52.8 mAP. As shown in Table II, following the evaluation metrics provided in [13], our model is evaluated on three different HOI categories, namely full, rare, and non-rare with default settings. The full proposed method acquires a steady 20.31 mAP on the HICO-DET test dataset, which could attribute to representative body-part features and potential relationship features in the scene.

D. Qualitative Results

Fig. 6 shows the qualitative outcomes and compares the HOIs detection results of our final (blue scores) and baseline (yellow scores) models. The representative human-object pairs in these images contain variance in object size, human body sizes, and different interaction categories. The interaction prediction probabilities of the correct interaction are visualized. For difficult HOIs, we observe that our final approach yields more dependable results.

Even when the crowd is dense and the spatial distribution of the crowd is uniform, or, the interaction is subtle and the object is tiny and interacts with some representative body part, SMPNet improves the score based on the baseline performs well. This shows that additional information is provided for these categories of interaction from visual features at the part and scene levels.

Special cases: Fig. 7 illustrates our proposed method's success and failure cases in multi-person and multi-object scenes. In Fig. 7(a) and Fig. 7(b), the interactive objects are the same categories. Due to the suppression of representative key parts and scene spatial information, in these two figures, (1,2) and (3,4) human-object pairs have obtained high values, and the values of the (1,4) and (3,2) human-object pairs are all approaching 0. However, when the visual or spatial overlap is high, and the representative parts and the spatial information of the scene are both confusing, the proposed method also products to erroneous predictions. Fig. 7(c) corresponds to the (3,6),(5,8), (5,4), and (7,6) human-object pairs in Table III, and Fig. 7(d) corresponds to the (1,4) and (2,3) human-object pairs in Table IV have obtained high values. At the same time, due to the suppression of representative key parts and scene spatial information, the values of (5, 4) and (3, 6) human-object pairs in Table IV are suppressed.

TABLE I. Performance Comparison on the V-COCO [5] Dataset. The Most Competitive Methods in Each Category Dataset Are Shown in Bold

Method	Feature Backbone	mAP_{role}
Gupta et al. [5]	ResNet-50-FPN	31.8
InteractNet [6]	ResNet-50-FPN	40.0
Kolesnikov et al. [43]	ResNet-50	41.0
GPNN [40]	Deformable ConvNets [48]	44.0
iCAN [7]	ResNet-50	45.3
Wang et al. [8]	ResNet-50	47.3
RPNN [44]	ResNet-50	47.5
Li et al. [45]	ResNet-50-FPN	47.8
Zhou et al. [46]	ResNet-50	48.9
PMFNet [47]	ResNet-50-FPN	52.0
Our baseline	ResNet-50-FPN	49.3
Our method	ResNet-50-FPN	**52.8**

TABLE II. The HOI Detection Performance on the HICO-DET [13] Test Set With the Default Setting (MAP×100). The Most Competitive Methods in Each Category Dataset Are Shown in Bold

Method	Feature Backbone	Default		
		Full	Rare	Non-rare
Shen et al. [49]	VGG-19	6.46	4.24	7.12
InteractNet [6]	ResNet-50-FPN	9.94	7.16	10.77
GPNN [40]	Deformable ConvNets [48]	13.11	9.34	14.23
iCAN [7]	ResNet-50	14.84	10.45	16.15
Wang et al. [8]	ResNet-50	16.24	11.16	17.75
RPNN [44]	ResNet-50	17.35	12.78	18.71
PMFNet [47]	ResNet-50-FPN	17.46	15.65	18.00
Our baseline	ResNet-50-FPN	15.68	12.82	16.54
Our method	ResNet-50-FPN	**20.31**	**17.14**	**21.26**

throw baseball 0.18, 0.83 **work on laptop** 0.32, 0.93 **read book** 0.41, 0.89 **hold cup** 0.14, 0.67

surf surfboard 0.75, 0.91 **talk on phone** 0.51, 0.90 **jump skateboard** 0.60, 0.82 **ride bicycle** 0.82, 0.96

eat pizza 0.34, 0.64 **catch frisbee** 0.28, 0.87 **kick soccer** 0.21, 0.81 **cut knife** 0.24, 0.89
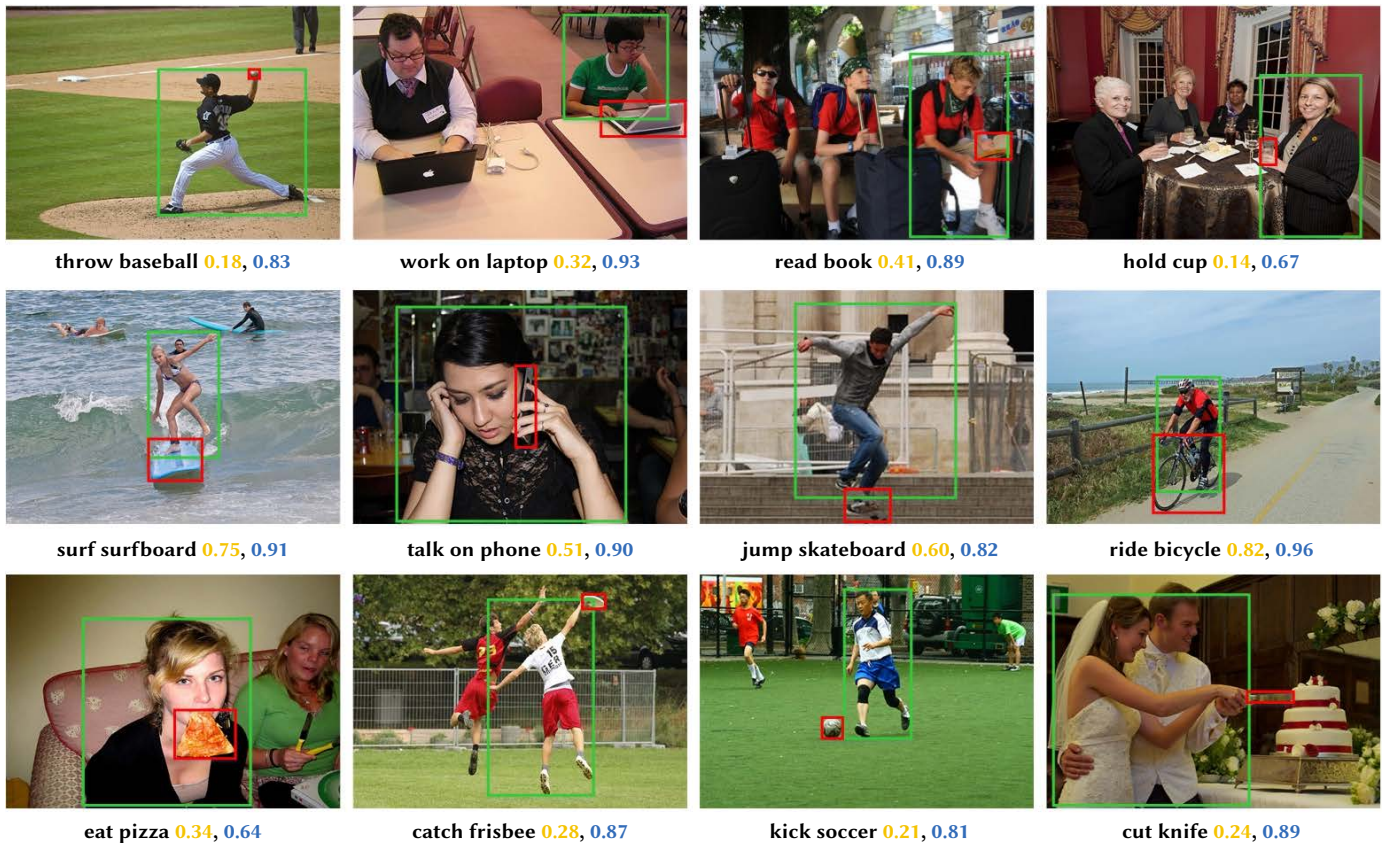
Fig. 6. The qualitative output of the final and baseline approaches on the V-COCO [5] test set. Yellow values and blue values denote scores predicted by the base model (instance level visual branch only) and SMPNet, respectively.

(a)　　(b)

(c)　　(d)
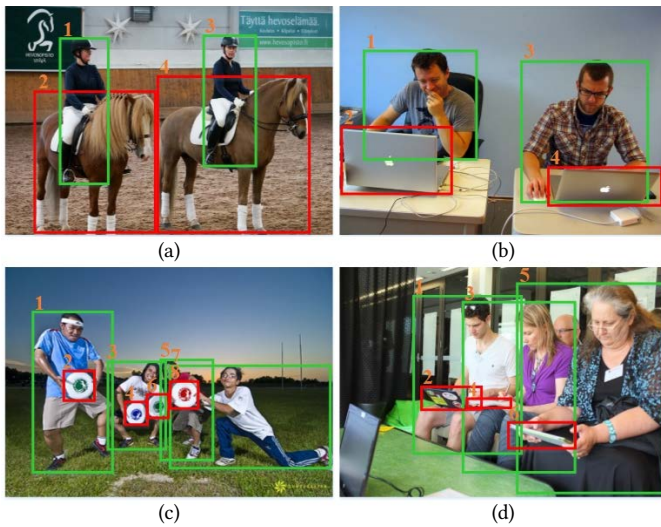
Fig. 7. The success(a and b) and failure(c and d) results of our proposed method for HOI detection. The scores of human-object pairs in failure results (c) and (d) correspond to Tables III and IV, respectively.

TABLE III. The "Hold Frisbee" Interaction Score of Human-Object Pairs in Fig. 7(C)

Instance number	1	3	5	7
2	0.8542	0.0017	0.0009	0.0002
4	0.0023	0.5621	0.3821	0.0042
6	0.0008	0.5308	0.5025	0.3831
8	0.0006	0.0026	0.4499	0.7259

TABLE IV. The "Look" Interaction Score of Human-Object Pairs in Fig. 7(D)

Instance number	1	3	5
2	0.6725	0.3899	0.0008
4	0.4335	0.8214	0.0026
6	0.0015	0.0056	0.8572

E. Ablation Studies

In this section, we conduct ablation study experiments using the V-COCO dataset to assess the efficiency of the different components of the proposed model. As formerly mentioned, we consider the basic model as an instance-level visual branch without part and scene levels, as in [7].

Spatial attention scores. We refer to PLVB when a variant of the part-level visual branch is analyzed without the spatial attention scores from the spatial configuration map. This branch does not use refined pose-part area features but directly utilizes the features obtained by the ROIAlign algorithm on shared feature maps for concatenation operation. We call the branch network using spatial attention scores as S-PLVB. As shown in Table V, spatial attention scores enhance HOI detection ability by 0.4 mAP.

TABLE V. Results of Ablation Studies on the V-COCO Dataset

Model	mAP_{role}
Baseline	49.3
Baseline+PLVB	50.7
Baseline+S-PLVB	51.1
Baseline+SGB	50.9
Baseline+S-PLVB+SGB	52.5
Our method (Baseline+S-PLVB+SGB+RS)	52.8

Part-level visual branch with spatial attention (S-PLVB). This is the vital component. Enlarging and capturing the features of the human pose-part area can effectively obtain relevant information of representative critical parts in interactions. A variation of our model is constructed to assess the impact of this branch. In comparison with the results of the basic model, the mAP utilizing S-PLVB is significantly improved from 49.3 mAP to 51.1 mAP, as shown in Table V.

Spatial graph branch (SGB). In this model branch, we construct a bipartite graph simulating scene with instances as nodes and obtain multiple human-object pairs relationship features in the whole scene through it. A variant of the model we proposed is executed without this branch. Compared with other results, the experiment shows an improvement of 1.6 mAP, as shown in Table V.

Relationship score (RS). Similarly to [47], we utilize the s_r to estimate the existence of an interactive relationship between humans and objects. Its purpose is to inhibit the score value without an interactive relationship. We state that the relationship score is based on the entire existence of the S-PLVB and SGB. In other situations, we straight fuse the results features of the three branches. The RS enhances performance by 0.3 mAP, as shown in Table V.

V. Conclusion

In our research, we proposed an effective human-object interaction detection model SMPNet, that utilizes a multi-level feature parsing strategy. It uses the instance, part, and scene levels parsing branches. In addition to the usual instance-level visual features, we introduce the pose of each interaction instance and the features of the keypoints' region, and utilize the spatial configuration map to generate spatial attention features to refine the visual features of these two levels. In the scene-level branch, we use graph neural networks to simulate the interaction between pairs in the entire scene, and add the spatial information of human-object pairs to adjust visual features. Finally, using the V-COCO and HICO-DET datasets, we demonstrate that our proposed model greatly increases detection capability and exceeds state-of-the-art techniques.

Acknowledgment

References

[1] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, G. Cottrell, "Understanding convolution for semantic segmentation," in *2018 IEEE winter conference on applications of computer vision (WACV)*, 2018, pp. 1451– 1460, IEEE.

[2] J. Lu, M. Nguyen, W. Q. Yan, "Deep learning methods for human behavior recognition," in *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2020, pp. 1–6, IEEE.

[3] L. Mi, Z. Chen, "Hierarchical graph attention network for visual relationship detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13886–13895.

[4] A. Gupta, A. Kembhavi, L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 10, pp. 1775–1789, 2009.

[5] S. Gupta, J. Malik, "Visual semantic role labeling," *arXiv preprint arXiv:1505.04474*, 2015.

[6] G. Gkioxari, R. Girshick, P. Dollár, K. He, "Detecting and recognizing human-object interactions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8359–8367.

[7] C. Gao, Y. Zou, J.-B. Huang, "ican: Instance- centric attention network for human-object interaction detection," *arXiv preprint arXiv:1808.10437*, 2018.

[8] T. Wang, R. M. Anwer, M. H. Khan, F. S. Khan, Y. Pang, L. Shao, J. Laaksonen, "Deep contextual attention for human-object interaction detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5694–5702.

[9] A. Bansal, S. S. Rambhatla, A. Shrivastava, R. Chellappa, "Spatial priming for detecting human-object interactions," *arXiv preprint arXiv:2004.04851*, 2020.

[10] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, J. Feng, "Ppdm: Parallel point detection and matching for real-time human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 482–490.

[11] S. Ren, K. He, R. Girshick, J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[12] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7103– 7112.

[13] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, J. Deng, "Learning to detect human-object interactions," in *2018 ieee winter conference on applications of computer vision (wacv)*, 2018, pp. 381–389, IEEE.

[14] V. Mnih, N. Heess, A. Graves, *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[15] R. Yu, K. Yang, B. Guo, "The interaction graph auto-encoder network based on topology-aware for transferable recommendation," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 2403–2412.

[16] R. Yu, B. Guo, K. Yang, "Selective prototype network for few-shot metal surface defect segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–10, 2022.

[17] B. Guo, Y. Wang, S. Zhen, R. Yu, Z. Su, "Speed: Semantic prior and extremely efficient dilated convolution network for real-time metal surface defects detection," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 12, pp. 11380-11390, 2023.

[18] D. Bahdanau, K. Cho, Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Advances in neural information processing systems," *Proceedings of Machine Learning Research*, pp. 5998–6008, 2017.

[20] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, 2005, pp. 886–893, IEEE.

[21] S. R. Sain, "The nature of statistical learning theory," *Technometrics*, vol. 38, no. 4, pp. 409, 1996.

[22] Y. Freund, R. E. Schapire, *et al.*, "Experiments with a new boosting algorithm," in *icml*, vol. 96, 1996, pp. 148– 156, Citeseer.

[23] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.

[24] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[25] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You only look once: Unified, real-time object detection," 2015, https://doi.org/10.48550/arXiv.1506.02640.

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, A. C. Berg, "SSD: Single shot multibox detector," *Lecture Notes in Computer Science*, vol 9905, pp 21–37, 2016.

[27] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, L. Fei-Fei, "Image retrieval using scene graphs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3668–3678.

[28] J. Yang, J. Lu, S. Lee, D. Batra, D. Parikh, "Graph r- cnn for scene graph generation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 670–685.

[29] H. Zhang, Z. Kyaw, S.-F. Chang, T.-S. Chua, "Visual translation embedding network for visual relation detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5532– 5540.

[30] B. Xu, Y. Wong, J. Li, Q. Zhao, M. S. Kankanhalli, "Learning to detect human-object interactions with knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2019-2028.

[31] S. Wang, K.-H. Yap, J. Yuan, Y.-P. Tan, "Discovering human interactions with novel objects via zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11652–11661.

[32] A. Bansal, S. S. Rambhatla, A. Shrivastava, R. Chellappa, "Detecting human-object interactions via functional generalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 10460–10469.

[33] S. Gao, H. Wang, J. Song, F. Xu, F. Zou, "An improved human-object interaction detection network," in *2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID)*, 2019, pp. 192–196, IEEE.

[34] 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.

[35] K. He, G. Gkioxari, P. Dollár, R. Girshick, "Mask r- cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[36] Z. Su, Y. Wang, Q. Xie, R. Yu, "Pose graph parsing network for human-object interaction detection," *Neurocomputing*, vol. 476, pp. 53-62, 2022.

[37] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, C. Lu, "Transferable interactiveness knowledge for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3585–3594.

[38] T. Gupta, A. Schwing, D. Hoiem, "No-frills human- object interaction detection: Factorization, layout encodings, and training techniques," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9677–9685.

[39] L. Li, Z. Gan, Y. Cheng, J. Liu, "Relation-aware graph attention network for visual question answering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10313–10322.

[40] S. Qi, W. Wang, B. Jia, J. Shen, S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 401–417.

[41] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014, pp. 740–755, Springer.

[42] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117– 2125.

[43] A. Kolesnikov, A. Kuznetsova, C. Lampert, V. Ferrari, "Detecting visual relationships using box attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 1749-1753.

[44] P. Zhou, M. Chi, "Relation parsing neural network for human-object interaction detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 843–851.

[45] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, C. Lu, "Transferable interactiveness knowledge for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3585–3594.

[46] T. Zhou, W. Wang, S. Qi, H. Ling, J. Shen, "Cascaded human-object interaction recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4263–4272.

[47] B. Wan, D. Zhou, Y. Liu, R. Li, X. He, "Pose- aware multi-level feature network for human object interaction detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9469–9478.

[48] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[49] L. Shen, S. Yeung, J. Hoffman, G. Mori, L. Fei-Fei, "Scaling human-object interaction recognition through zero-shot learning," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 1568–1576, IEEE.

Zhan Su

Zhan Su received his B.S. and M.S. degree in software engineering from Northeastern University, Shenyang, China, in 2015 and 2017, respectively. He is currently a Ph.D. candidate at Northeastern University, Shenyang, China. His research interests include computer vision, machine learning, and action detection.

Ruiyun Yu

Ruiyun Yu is currently a professor and vice dean of the Software College at the Northeastern University, China. He received his Ph.D. and M.S. degree in computer science and bachelor degree in Mechanical Engineering from the Northeastern University in 2009, 2004, and 1997, respectively. He serves as the director of center for Cross-media Artificial Intelligence. He is one of the Baiqianwan Talents of Liaoning Province, China (Hundred Talents Level), and now a member of the CCF IoT Committee, and a Senior Member of CCF. His research interests include intelligent sensing and computing, computer vision, data intelligence, etc.

Shihao Zou

Shihao Zou received the B.Sc. degree from Beijing Institute of Technology, China, in 2017, and the M.Res. degree from University College London, UK, in 2018. He is currently a Ph.D. candidate at University of Alberta. His interests include computer vision and machine learning, especially human pose and shape estimation, motion capture system.

Bingyang Guo

Bingyang Guo is currently a PhD candidate in the software college of Northeastern University China. He received his Bachelor degree in mechanical engineering from Shenyang Ligong University, China, in 2018, and Master degree in mechanical design and theory from Northeastern University, China, in 2021. His research focuses on image segmentation, image restoration and defect detection.

Li Cheng

Li Cheng received the Ph.D. degree in computer science from the University of Alberta, Canada. He is an associate professor with the Department of Electrical and Computer Engineering, University of Alberta. He has previously worked at A*STAR, Singapore, TTI-Chicago, USA, and NICTA, Australia. His research expertise is mainly on computer vision and machine learning. He is a senior member of the IEEE.

Aligning Figurative Paintings With Their Sources for Semantic Interpretation

Sinem Aslan[1,2], Luc Steels[3]

[1] Ege University, International Computer Institute, Bornova, Izmir (Turkey)
[2] Ca' Foscari University of Venice, DAIS & ECLT, Venice (Italy)
[3] Barcelona Supercomputing Center, Barcelona (Spain)

* Corresponding author. sinem.aslan@unive.it

uniR
LA UNIVERSIDAD
EN INTERNET

Abstract

This paper reports steps in probing the artistic methods of figurative painters through computational algorithms. We explore a comparative method that investigates the relation between the source of a painting, typically a photograph or an earlier painting, and the painting itself. A first crucial step in this process is to find the source and to crop, standardize and align it to the painting so that a comparison becomes possible. The next step is to apply different low-level algorithms to construct difference maps for color, edges, texture, brightness, etc. From this basis, various subsequent operations become possible to detect and compare features of the image, such as facial action units and the emotions they signify. This paper demonstrates a pipeline we have built and tested using paintings by a renowned contemporary painter Luc Tuymans. We focus in this paper particularly on the alignment process, on edge difference maps, and on the utility of the comparative method for bringing out the semantic significance of a painting.

Keywords

I. The Comparative Method

THIS paper reports on research into the artistic methods used by figurative painters using computational algorithms. One aspect of the artistic method concerns style. Human viewers quickly see whether a landscape or a face is painted in a romantic, impressionist, expressionist or cubist style. Much work in AI, with remarkable results, has been done on capturing an artist's style or period and generating new works in a similar style [1], [2]. Although style transfer is very interesting, it is not discussed in this paper because we are interested in another aspect of the artistic method, namely the *expression of meaning*.

Painters, particularly figurative painters, want to mean something with their work and they introduce *signifiers* that convey these meanings. They introduce focal points and centers of interest, pronounced edges, textural regions with less detail, brightness contrasts, unusual defigurations of objects, etc. [3]. Whether these potential signifiers become real signifiers is determined by later semantic processing that make use of the context and world knowledge to interpret the painting. For example, a portrait is typically not a photographic rendering of the depicted person's face. The painter selects, highlights, and transforms the source image (either a live model or a photograph) in order to express meanings at many levels. For example, he or she may want to convey the personality and affective state of the person, his or her moral attitudes, and the socio-

cultural context in which the person lived. This is well illustrated by Francis Bacon's series of popes all inspired by the famous painting by Velasquez of Pope Innocent X.

There is already a vast and fast growing literature using computer vision algorithms for various functions related to art interpretation, such as classification [4], object detection [5], similarity retrieval [6], sentiment analysis [7], and generative art [8] to name just a few of the most active areas. Most of this work is based on the use of low level image analysis and image transformation although with the advent of deep learning and the availability of semantically annotated image datasets, there is now a general trend towards the detection and interpretation of features that contribute to meaning [9].

Using this prior art, there are two approaches to study the artistic methods that artists use to transform source images into paintings. One approach is to start from a photograph of a face or a real world situation like a still life, construct a semantic interpretation that includes various levels of meaning such as affect, character, perspective, moral implication, socio-cultural context, etc. and then convert this photograph plus its desired meanings into a painting by making informed choices about cropping, lighting, color and tone, brush strokes, edges, level of textural detail, contrast, etc. [10]. This approach is being pursued in the context of *non-photorealistic* or *artistic rendering* (*NPR*) [11], which has become more and more sophisticated lately to

include parameters driven by semantic criteria such as emotional or personality analysis rather than random perturbations [12].

Another complementary approach, advocated in this paper, goes in the other direction. We call it the *comparative method*. It starts from the source photograph and background knowledge about the painter and from the catalog and studies how this painter actually transformed the source photograph and what possible meanings could have played a role. This approach is therefore a way to study a painter's artistic method, not just his style, as is done in research on style transfer, but what kind of meanings have been found to be important to express and what expression strategies have been employed. The benefit of a comparative method is not only increased understanding of a painting and the oeuvre and approach of a painter. This kind of analysis could also yield insights and methods that could feed into the NPR approach by shedding more light on how creative artists achieve non-photorealistic rendering. Normally this method can only be used if a source image (which could also be an earlier painting by the same or another painter) is available, however there have also been remarkable experiments, in relation to the work of Rembrandt, where a new photograph is made of an existing person that resembles a figure painted centuries ago [13] and then we can use the comparative method starting from this photograph.

This paper takes steps toward an application of the comparative method. It requires that we find first the source of the painting under investigation by interacting directly with the painter or by historical research. The massive number of images now available on the Internet and existing image search algorithms should be very helpful in this respect. The next step is to overlay the relevant parts of the source on the target painting so that they become visually comparable. Painters may isolate only a small area of a source. They may leave out details, for example to make the object on the painting less tied to its source and hence more universal. They may stretch represented objects, shift them with respect to each other and change their orientation. They may change the color choices of significant surfaces, add or remove edges, etc. Many of these actions are geared towards the creation of potential signifiers and influence the way a viewer reacts to the painting.

More concretely, we focus first in this paper, which builds further on the results reported in [14], on two concrete technical challenges necessary to make the comparative method applicable: (i) finding the geometric operations of translation, scaling and rotation which align the painting and its source, and (ii) computing edge difference maps. In a final section we go back to the bigger picture and demonstrate the utility of difference maps, specifically by extracting facial activation units and focusing on those where the edge difference maps have identified regions of interest.

II. The Case Studies

Using paintings by Luc Tuymans, a contemporary Flemish painter, we have done a number of concrete case studies for both technical challenges. It's worth to note that working with a living artist makes it possible to validate our methodology and tells us whether the algorithms have yielded valuable results, not only for viewers, curators or art historians but also for those who create the artworks. These case studies have lead to an exhibition called 'Secrets'[1], Artificial Intelligence and Luc Tuymans' at the BOZAR cultural center in Brussels, which raised considerable impact in the community of artists, art curators and interested viewers.

Luc Tuymans is currently considered one of the most important contemporary painters [15]. His solo exhibitions took place at some of

[1] https://readymag.com/u3083945729/secrets-guide/

the most prestigious and influential art centres in the world, such as the MOMA in New York, the MCA Museum of Contemporary Art in Chicago, the Palazzo Grassi in Venice, the Städel Museum in Frankfurt, the National Art Museum of Beijing, BOZAR in Brussels, etc. We have been fortunate to be in direct and frequent contact with this painter and to have access to relevant parts of his digital archives. Luc Tuymans is very articulate in describing both his own artistic method and the methods used by other painters [16] and he almost always works on the basis of a photographic image, which he has supplied or validated for the case studies we have undertaken.

We have examined quite a number of paintings from the 2019-2020 solo exhibition of Luc Tuymans in the Palazzo Grassi in Venice. In this paper we demonstrate our work using two oil paintings from this exhibition, shown in Fig. 1: *K.*, which depicts the face of a young woman looking energetically into the future, and *Secrets*, which depicts the face of an older man in a somber mood. *K.* is bigger than *Secrets* and they are quite different in terms of color usage and general emotional impact.
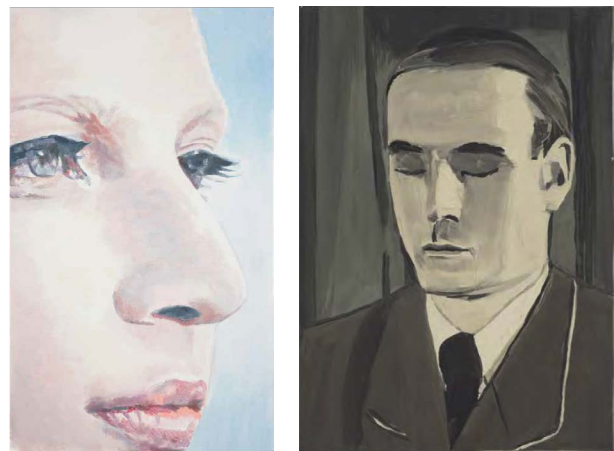
Fig. 1. Two of the oil paintings by Luc Tuymans used in our case studies. Left: *K.* (2017) oil on canvas 135 x 80,2 cm. Private collection. Right: *Secrets* (1990) oil on canvas 52 x 37cm. Private collection.

The overall workflow we used for this paper is illustrated for the painting *K.* in Fig. 2. From left to right, there is the identification of the original, the alignment process, edge detection and construction of comparing edge maps, and their use in further pattern recognition and semantic interpretation. Each of these steps is discussed in detail in the body of this paper, both for the painting *K.* and for *Secrets*.

III. Sources

Due to direct interactions with the painter, we had access to the originals he used. But it is also interesting to try and find the sources of these originals and their wider context using the Internet. We used reverse image search as offered by several commercial search providers, namely Google (American), TinEye (American), Bing (American), Yandex (Russian) and Baidu (Chinese), using the paintings themselves as the key. These search engines indeed provide a large number of images that are visually related to the painting, with interesting cultural differences between the search engines, undoubtly based on the image repositories used to train the reverse image search algorithms. However, none of them yielded the original photograph nor the context in which it was taken. We hypothesize that this difficulty is related to the well known domain adaptation problem: Two images that humans see as quite similar are nevertheless not detected to be so due to differences in illumination, pose, image quality, texture, etc., because they cause a distribution change or domain shift between the domains
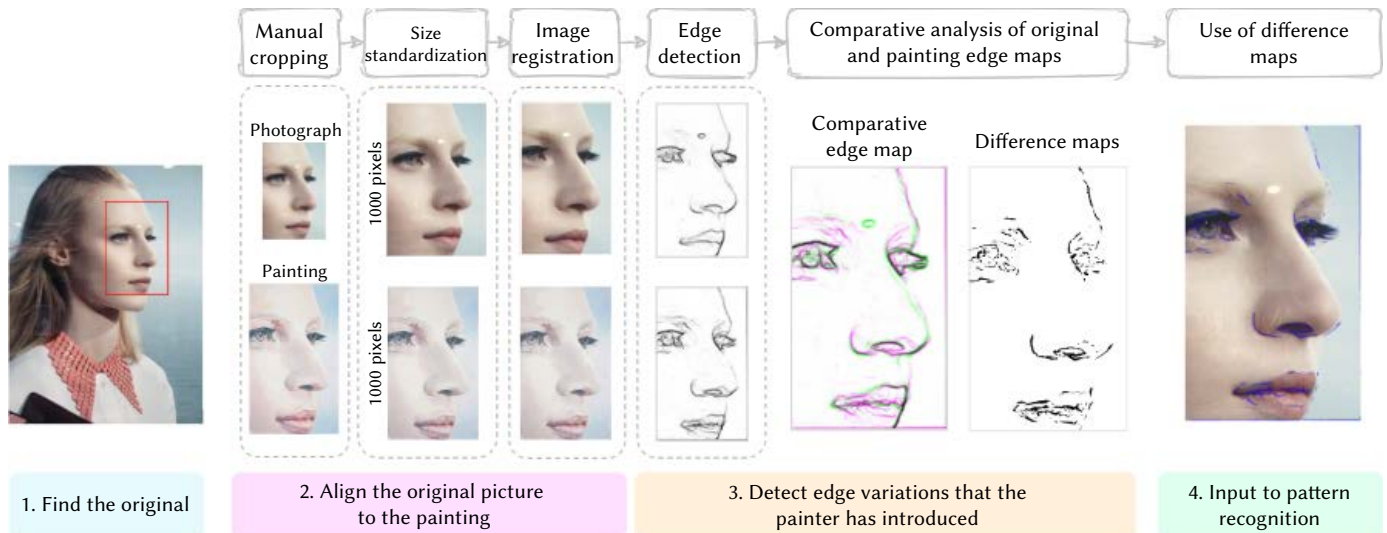
Fig. 2. Work flow discussed in this paper, going from identification and alignment to the construction of edge difference maps. It is illustrated for the painting *K*.

derived from the respective images [17]. Solving domain adaptation is currently a frontier area in computer vision and we may expect that search engines will get better if new results are incorporated.

On the other hand, when we provide more information to search engines, we can retrieve the originals and their context. More information means that we supply the title or text from the catalog or we supply the source image that the painter originally provided to us. For the painting K. we show the results in Fig. 3. The painting is in fact a close-up of a face which is itself part of a larger scene coming from a Dior commercial. Knowing this context suggests that the direct inspiration is a fashion model from an advertising campaign. We see an objectification [18] of the human body, more concretely in this case of the human face, which is typical for advertising imagery in fashion or cosmetics. It can be said that this objectification is present in the photograph, and even more so in the painting considering the followings: with an excessive focus on the face, the context was almost completely eliminated, the details that normally make the faces look alive were softened, and the letter K. was chosen for the title of the painting instead of a real name for the woman depicted. Marc Donnadieu, curator of the Palazzo Grassi exhibition points in the catalogue to additional features of the face: 'smiles discreetly', 'defiant', 'expressive gaze', which are evoked through signifiers such as subtle changes in the lips, eyes and eyebrows, and a change in the nose.

Fig. 3. From left to right: painting, artist supplied source, original image from the Dior Autumn-Winter 2015 campaign photographed by Willy Vanderperre; clothes designed by Raf Simons and the fashion model is Julia Nobis.

For the painting *Secrets*, the original source is shown in Fig. 4. It is in fact a famous photograph of the Nazi architect and Third Reich minister of armament Albert Speer. The photograph was taken by Walter Frenz, the chief cameraman of Leni Riefenstahl. Seeing this source makes us realize at once that the secrets mentioned in the title have to do with denying knowledge and responsibility for the atrocities of the war. The painter has used again the technique

of zooming in on a segment of the original image and on removing iconic signifiers in the source (such as the Nazi insigna) in order to make the portrait more timeless and convey expressive and emotional meanings. We see also that the face has become more rectangular, almost looking like a mask. The eyes are closed signifying the hiding of secrets, a grey shadow hangs over the face, the nose is sharper, and the lips are tight.

Fig. 4. From left to right: painting, cropped original, original. The cropped original is a photograph of Albert Speer provided by the artist. The original has been found through Google search using 'Albert Speer' as the key. It did not appear through reverse image search, neither with the painting as key nor with the cropped image supplied by the painter as key.

IV. Aligning the Source Image to the Painting

We first focus on the geometric transformation process. We need algorithms that compute how the source was transformed to obtain the target painting, in other words the transformations that allow the source to be aligned as much as possible with the painting. Subsequent comparative visual processing rely entirely on whether such an alignment could be established. The relevant technique from computer vision for this purpose is called *image registration* now also often called *image alignment*. Image alignment is a well-studied problem in image processing and ready-made algorithms are available of all common computer vision platforms such as Matlab or OpenCV. We used the image registration algorithms available on Matlab².

Alignment aims to spatially align multiple images of the same scene. Image alignment is widely used in a variety of application fields [19] for: (1) multi-view analysis, where images from different viewpoints of the same scene are aligned for a larger (either 2D or 3D) representation of the scene; (2) multi-temporal analysis, where images of the same

---

² https://mathworks.com/discovery/image-registration.html

scene taken at different times are aligned to detect changes over that time period; or (3) multi-modal analysis, where images of the same scene are acquired by different types of sensors and aligned for fusing information from different sources to obtain a more comprehensive representation of the scene [19].

For successful registration some pre-processing of the image data was required. If a painting encompasses only a smaller portion of a picture (as is the case here), the relevant region in the picture has to be cropped to exclude the uninterested region and automatic image alignment has to be performed on the cropped portion. Performing cropping before image alignment has been recommended as a useful operation in the literature. For example, following a feature-based image registration approach [20] demonstrated that cropping the overlapping area and thus restricting the search area generally improves the results of the feature matching algorithms. [21] showed that cropping images by discarding the unnecessary background aids the alignment algorithm because it ensures registration of the foreground rather than registration of the background which improves the registration performance.

Moreover, as mentioned in [22], [23] the cropping operation reduces registration time and prevents memory issues. While cropping can be done manually [20]–[22], [24]–[26], there have been a number of experiments to automate it [23], [27]. Here we do not investigate automated cropping but use hand-selected possible candidates for *K.* and *Secrets* as shown in Fig. 5. The manual selection does not need to be very precise but without it we do not get usable results, as shown in Fig. 8.
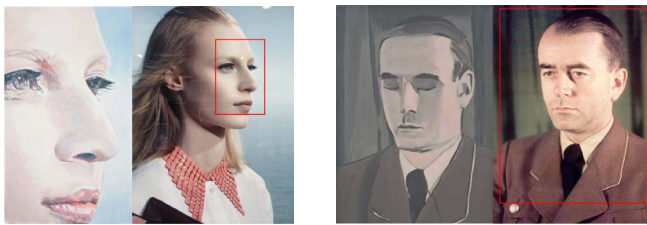
Fig. 5. Painting and selected source image for cropping *K.* (left) and *Secrets* (right).

In the terminology of image alignment algorithms, there are two images given as input. The one that remains unchanged is called the *reference* or *fixed* image, while the other, that is transformed to align with the reference image, is called the *sensed* or *moving* image [19], [28]. In this work, we consider the photograph, which is the original source image, as the *moving image* and try to align it to the painting by successively applying geometrical transformations on it. Thus, the *painting* is considered as the *fixed image* in image alignment terminology. This obviously reflects the *perspective of the painter* as the painter transforms the original photograph into a painting. Thus, in this paper, we discuss the painter's perspective.

[19] summarise the main steps of the majority of alignment methods as follows: 1) Either use *feature detection* to find salient elements such as lines, keypoints, or regions in both images, which can be done by well-known algorithms such as MSER, Harris, SURF, SIFT, etc., or use image pixels densely sampled on a regular grid as features [29]; 2) *Match* these features between the two images by means of similarity or correlation of the local neighborhoods of the features; and 3) *Estimate a transform model* to find the mapping function that transforms the moving image so that it overlays as well as possible with the reference one. We now show the result of applying these steps for our case study.

A. Feature Detection and Matching

For these phases of the image alignment process, feature- based and area-based methods can both be used [19] and we have tried known

algorithms for each method in order to see which approach is the most appropriate in the present context.

Feature-based approaches aim to match detected salient structures in both reference and moving images. We used the well known SURF algorithm on the gray-scaled original photograph and the painting [30]. We then *l*2-normalized the feature vectors to obtain the unit vectors and matched the features in the original photograph to the nearest neighbors in the features of the painting by computing the pair-wise distances (sum of squared differences) between feature vectors in the photograph and the painting. We used the default highest value for the match threshold T ($0 < T \leq 100$) of the software platform[3], which is $T = 100$, i.e., two feature vectors match when the distance between them is less than or equal to 100. We performed a forward and backward matching between the photograph and the painting, and kept the best matches of the feature vectors. Results are shown in Fig. 6. For *K.*, 47 and 197 features were detected in the photograph and the painting, respectively and only six of them matched, while for *Secrets* 134 and 249 features were detected in the photograph and the painting, respectively, and 17 of them matched.

For a successful image alignment, the number of correctly matched features between the reference (fixed) and moving (sensed) image should be sufficiently high regardless of the geometrical or photometrical changes in the images [19]. This is not the case here. We observe in Fig. 6 that only two of the matches can be accepted as correct in the case of *K.*. On the other hand, while the number of matched points is higher, compared to *K.*, for *Secrets* there are still an unreasonable number of mismatches and they significantly affect alignment accuracy. We obtained similar results for other feature-based alignment algorithms, specifically the well known SIFT [31] and ORB [32] algorithms.
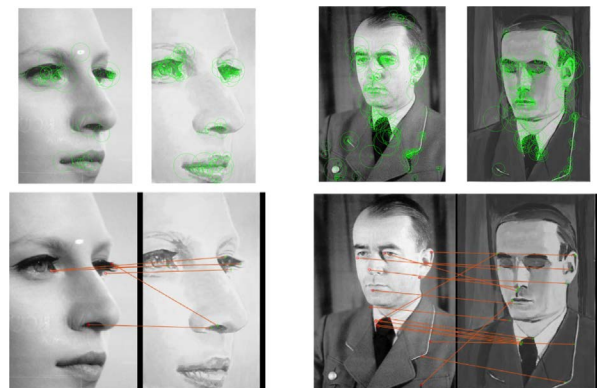
Fig. 6. Top: Detected SURF keypoints in the cropped photograph and the painting of *K.* (number of detected keypoints for the original source and painting images is 47 and 197, respectively), and *Secrets* (number of detected keyponts for the original source and painting images is 134 and 249, respectively); Bottom: Matched SURF features for *K.* and *Secrets* (number of matched features for K. and Secrets is 6 and 17, respectively).

Once there is a feature correspondence, the mapping function, also called transform model, can be estimated. It transforms the moving image so that it overlays as well as possible with the reference image, which requires finding a transformation function and estimating its parameters. A transform model hence characterizes the geometrical deformation to which the moving image has been subjected by the painter. For the present study we restricted the possible transforms to be shape-preserving so there could only be *rotation*, *translation*, and *isotropic scaling*.

Following feature-based matching, the M-estimator SAmple

---

3 https://mathworks.com/help/vision/ref/matchfeatures.html

Consensus (MSAC) algorithm [33] was used to estimate the transform model parameters. The MSAC algorithm is a variant of the Random Sample Consensus (RANSAC) algorithm [33], [34], which is known to be more robust than RANSAC [35]. The quality of model estimation is evaluated using the sum of distances between all points to the estimated model differently from the RANSAC, which uses the number of inliers, i.e., correctly matched points, as the quality measure. Application of the transform to the moving image and overlay on the reference image is shown in Fig. 7. The results are not good at all, undoubtly because the feature-based approach used here (and for similar methods i.c. SIFT and ORB) does not work well for finding correspondences between photographs and paintings in the case of Luc Tuymans, simply because the operations done by the painter are too numerous - even though human observers immediately see that the same image contents are present. Perhaps a better result would be obtained if the painting is first transformed to look similar to a natural scene before alignment is attempted [36].
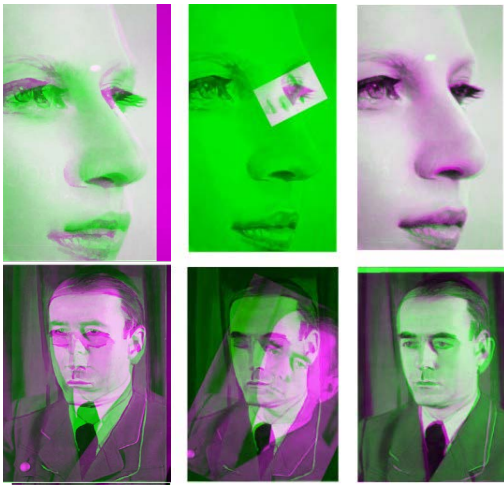
Fig. 7. Overlay of photograph on painting for *K.* (top) and *Secrets* (bottom): Gray regions in the composite image show where the two images have the same intensities. Magenta and green regions show where the intensities are different. Left: With the painting and the picture overlayed without image alignment; Middle: After image alignment using a feature-based (SURF) approach; Right: After image alignment using an area-based approach.

Area-based approaches do not attempt to detect salient regions, but use windows of predefined sizes or even entire images to estimate the correspondence. [19]. In this work, we used image pixel regions and one-plus-one evolutionary optimizer [37] for matching them, which is implemented using the Matlab Registration Estimator App[4] using its default parameter settings.

To speed up the process, this algorithm builds an image pyramid (both for the reference and moving input images) that has $N = 3$ levels where at each pyramid level the input image resolution is decreased by a factor of 2 in both image dimensions. Then, a coarse-to-fine hierarchical strategy is used to apply the alignment method, i.e., optimization starts at the coarsest level of the pyramid and continues at the finer levels until either convergence or *MaximumIterations* = 100 is reached. While going up to the finer resolutions, *estimates of feature correspondence* and *transform model parameters* are improved gradually [19]. The Mattes mutual information metric [38] is used to measure the similarity between reference and moving images in every optimization step. It was shown by [19] that this metric provides a more accurate alignment than the Mean Squares metric when the moving and reference images are from different modalities, as in our case.

[4] https://mathworks.com/help/images/register-images-using-the-registration-estimator-app.html

One-plus-one evolutionary optimizer [37] refines the estimation of the parameters for the specified (similarity) transform model iteratively. A set of variations, called the *children*, of a given matrix of transformation parameters, called *parent*, is initially created using aggressive perturbations. If a child's parameters bring a better alignment, it becomes the new parent on the next iteration, otherwise, the parent stays the same and new children are computed with less aggressive changes to the parent's matrix. At every iteration of the optimization, the moving image is resampled by bilinear interpolation based on the transformation model estimated in that step and the similarity to be optimized between the reference and the transformed moving image is computed. For example, the obtained transform model for *Secrets* includes a translation with $t_x = -37.12$ and $t_y = 28.21$, a scaling by a factor $s_x = s_y = 1.09$ and a slight rotation by $\theta = -0.515$ degrees.

Finally, using the estimated transform model (from the feature-based or area-based approach), the moving image is resampled by bilinear interpolation and thus the images are said to be registered. Image alignment results obtained by feature-based and area-based approaches are shown in Fig. 7. It is seen that the accuracy of the alignment results with the area-based approach provides us almost a perfect alignment and we can build further on that base. For the feature-based approach we see that the original *K.* image has been shrunk and overlayed on the left eye and the original *Secrets* image has been rotated, both of which are not usable for further analysis.

It is of course possible to try many other methods for image alignment. We just mention two other promising approaches: based on using mutual information (MI) as a metric for alignment, as used by [39] for example, or using point cloud representations as originally developed for the reconstruction of 3D objects from multiple sources (laser scanner, digital cameras), as illustrated in [40]. However the present result is adequate for the next step in the pipeline.
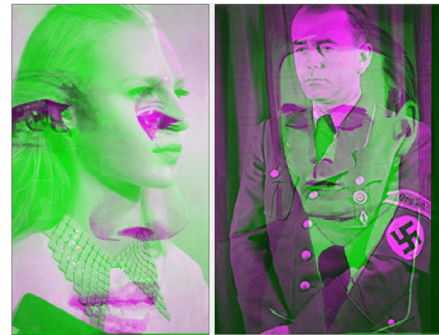
Fig. 8. Results of alignment using the same area-based approach as in Fig. 7 but now without cropping, both for *K.* (left) and *Secrets* (right). The algorithm establishes a transform model, which is however totally unsatisfactory for further comparison.

V. Edge Detection and Difference Maps

Having obtained an adequate alignment of the original picture with the painting, we can start to investigate the micro-transformations that the painter has introduced and their function [14]. These variations happen for different visual aspects, e.g., color, contrast, orientation, edges, contours, etc.. Thus, first these aspects are needed to be extracted from the painting and the aligned original picture, so that a comparison is possible. In the present work, we only look at edges, i.e., we explore which additional edges or edge variations the painter has introduced with respect to the original source photograph. Edges are certainly not the only vehicle that painters use to achieve visual effects and express meanings but it is a very important one. In our project

Detected edges		Binarized edges		Detected edges		Binarized edges	

| Source | Painting | Source | Painting | Source | Painting | Source | Painting |
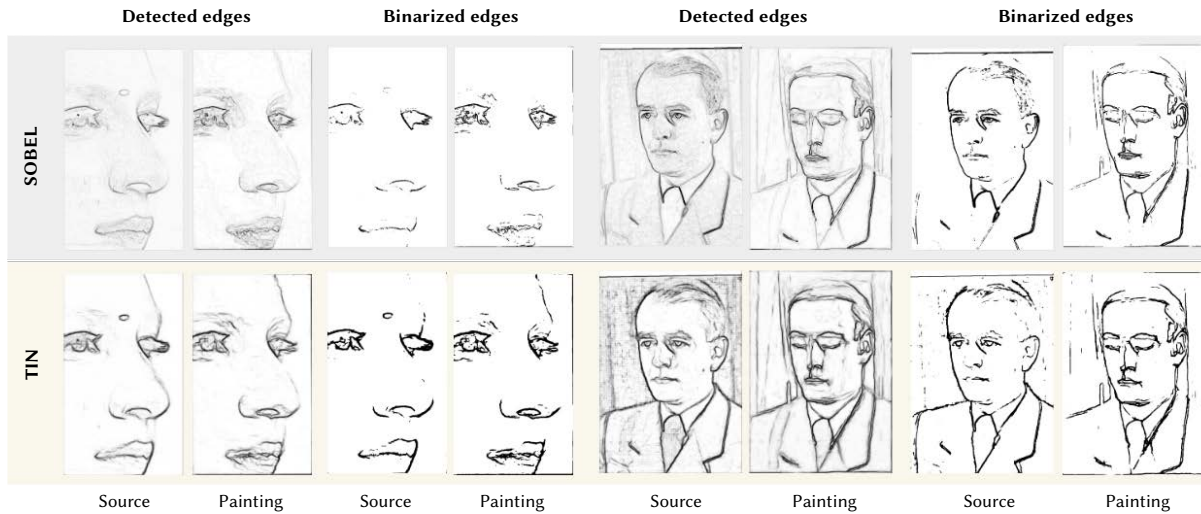
Fig. 9. Comparison of two edge detection methods, i.e., SOBEL [44] (1st row) and TIN [43] (2nd row), taking the painter's perspective (going from photograph to painting), with Non-Maximum Suppression (NMS) post-processing. We see that the TIN method provides clearer edges showing the variations introduced by the painter in a clearer way.

we have also been investigating other types of features, specifically as related to the focal point [41] and color [42] and of course we are considering other features as well.

Our methodology consists of two steps. First, we detect the edges in both the source image and the painting. Then, for a comparative analysis, we construct a *difference map* that show all the edges for the source picture (magenta) and the painting (green) simultaneously, a *similarity map* that shows only the shared edges, and a *difference map* that shows which edges do not exist in the source image.

A. Edge Detection

We wanted to compare the results with traditional edge detection methods, more specifically the *Sobel Isotropic* 3×3 *gradient operator* (SOBEL), and a deep neural network known as the *Traditional Inspired Network* (TIN) [43]. To achieve better-located edges, as mentioned in [43], we applied a post-processing operation, namely *Non-Maximum Suppression* (NMS), for both edge-detection methods. Specifically, we computed three edge maps for each input image at different scales, namely 1.5×, 1×, 0.5×, resized the resulting edge maps to the original image size, i.e. $x = 1000$, and averaged them to obtain the final edge map. More details about the two methods are as follows:

1. *The SOBEL edge detection* method [44], which has been widely used in image processing since the late 1960, is based on the derivation of a computationally efficient gradient operator. The gray-scaled input image I is convolved with the following 3×3 kernels, to obtain the gradients for horizontal and vertical intensity changes, that are G_x and G_y, respectively [44] as in Eq. (1).

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} * I, \; G_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * I \quad (1)$$

$$\tag{1}$$

Then, the resulting gradient approximations are combined with $G = \sqrt{G_x^2 + G_y^2}$ to have the gradient magnitude at each point in the image; this is the result displayed for the Sobel Method in the subsequent images in this paper.

2. *Deep neural networks* were originally designed for high-level computer vision tasks, e.g., object or scene recognition. Since edge detection is a simpler task, a lightweight deep learning network with reduced computational complexity can provide high-quality edges [43]. Motivated by this fact, in this work, we used a lightweight deep neural network architecture named *Traditional*

Method Inspired Network (*TIN*) [43] where state-of-the-art accuracy performances were reported on the BSDS500 test set.

The TIN framework is composed of three modules: *Feature Extractor*, *Enrichment*, and *Summarizer*, which roughly correspond to gradient, low pass filter, and pixel connection in the traditional edge detection schemes [43]. In particular:

- *Feature Extractor* is formed by the 3×3 convolutional neural network layer which is designed to simulate the gradient operators (such as Sobel operator).

- *Enrichment* aims to remove the noise or tiny/isolated edge candidates by using multi scale filtering dilated convolutions, and

- *Summarizer* produces the final edges by fusing the outputs of the previous layer.

Two architectures, TIN1 and TIN2, were proposed by [43]. TIN1 is composed of the aforementioned three modules and TIN2 is a stack of two TIN1s, where output of the first module of the first TIN1 is downsampled by max-pooling in half and given as input to the first module of the second TIN1. The pre-training of the TIN method was performed on three datasets, i.e., BSDS500 (natural images), PASCAL VOC (natural images), and NYUDv2 (indoor images) [43]. Since higher performances were reported by TIN2 compared to TIN1 in [43], we employed the TIN2 architecture using the code published by the authors5 and their pretrained model on the aforementioned datasets.

Once we computed the grayscale edge maps with either SOBEL or TIN, we binarized them so that the significant edges be highlighted more. In the binarization operation we used a global threshold computed using Otsu's method, which chooses the threshold to minimize the intra-class variance and accordingly maximizes the inter-class variance of the thresholded black and white pixels [45]. Fig. 9 shows edge maps computed by each method and the outcome after thresholding. After thresholding it is observed that a significant amount of edges detected by the SOBEL method were removed, while the higher quantity of edges detected by TIN method were preserved. We proceed for further analysis with the edges detected by the TIN method, since TIN preserved the significant edges better. Of course, it is possible to try still other edge detection algorithms, e.g., [46], and the Canny algorithm [47] would be a prime candidate, but given the results with TIN we continue with this solution.

---

5 https://github.com/jannctu/TIN

1. Comparative Analysis of Detected Edges

Finally, we compute the edge maps shown in Fig. 10. At the comparative edge map on the left side, detected grayscale edges by TIN are shown in magenta for the source picture, green for the painting, and black when the edges overlap. The equal edge map in the middle shows the overlapping edges in the binarized image, which can be considered as the locations where the painter has not introduced any modifications. We are most interested in the difference map on the right. It shows the edges introduced or emphasized by the painter and not found in the source image.
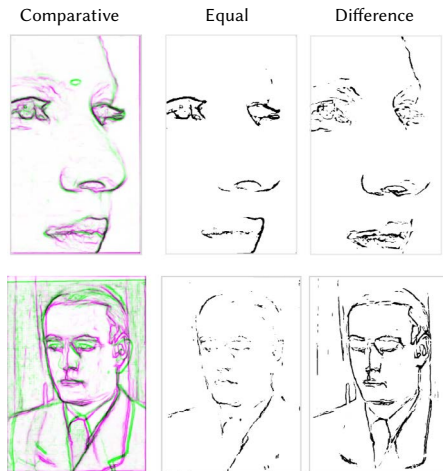
Fig. 10. Comparative analysis of the detected edges by TIN on the source image and painting. In the comparative grayscale edge map (left), the painting edges are in magenta and the source picture edges are in green. The similarity edge map (middle) shows the overlapped binarized edges. The difference edge map (right) shows the edge modifications that were introduced by the painter which do not exist in the source picture.

VI. Step Towards Semantic Interpretation

The edge difference maps are just one of the many difference maps we can make but instead of considering other difference maps, we turn to the topic of semantic interpretation which is the ultimate goal of our work. It is important to point out that probing for the presence and meaning of signifiers requires more than the alignment and low level feature analysis discussed so far. We need to apply pattern recognition, such as (i.e., speaking for portrait paintings) facial expression recognition [48], [49] or recognition of attributes such as glasses, lipstick, hat, gender, hair color and hair shape, eyebrow shape, nose shape, lip shape, race, face shape, existence and shape of moustache and/or beard, etc. [50], as well as knowledge level processing based on common sense, world knowledge and historical knowledge.

We discuss in this section first what we can potentially learn from the edge difference maps and next how edge difference maps can be used in cooperation with other pattern recognition algorithms, more specifically face detection, facial behavior analysis and emotion recognition to probe semantic interpretation. A discussion on the role of knowledge level processing is beyond the scope of this paper.

A. Semantic Interpretation Using Edge Difference Maps

Fig. 11 shows the edge difference map overlayed on the painting to illustrate that the following regions have been slightly altered [14]: (a) bottom part of the lips area, (b) the pupil and the area around the right eye, (c) the nose area and the curve at the right wing of the nose, (d) the left eye, especially the corner with the nose, and (e) the region above the left eye. These are therefore centers of interest and should be focal areas for subsequent pattern recognition and interpretation
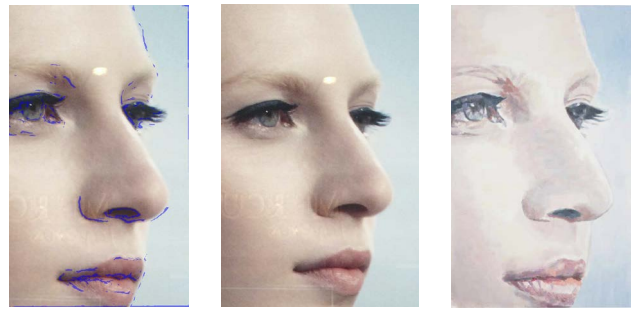
Fig. 11. Left: Original source image of *K.*. Middle: Edge difference map overlayed on aligned original source. Right: Painting *K.*

algorithms. The changes between the source and the painting are very subtle, nevertheless they give an overall change in the facial expression, as suggested by Marc Donnadieu in the exhibition catalogue: *K.* 'smiles discreetly', 'is defiant', and has an 'expressive gaze'. The eyes are more open towards the world on the painting, also because they are in a lighter blue. The right mouth corner emphasizes a faint smile.

Continuing the interpretation of *K.*, we observe that in addition to deleting all context and objectifying the woman in the image, there is a strong cropping of the original image (Fig. 5) which also causes a strong focus on the eye gaze and on the main components of the face: the eyes, the nose and the lips. This zooming and focusing is so strong that state-of-the-art image algorithms have great difficulty. For example, YOLOv3 [51] does not recognize that *K.* represents a face, but instead labels the eye area as a bird (Fig. 12) and Mask R-CNN, another common pixel labelling algorithm [52], does not do much better. Also state of the art facial behavioral analysis algorithms have difficulty to recognize the facial components. This is shown in Fig. 12 for the application of OpenFace [53]. For the painting *K.* (middle of Fig. 12), OpenFace 2.0 has difficulty to detect the facial landmarks, and recognize the facial action units and head orientation on the cropped source and it does not recognize any of them on the painting.
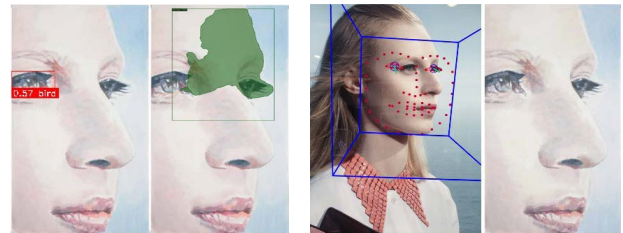
Fig. 12. Left: Segmentation and labeling of *K.* YOLO segments the eye and labels this segment as a bird with 0.57 % certainty and Mask R-CNN labels an area at the forehead as person with 0.74 % certainty. Right: Application of OpenFace to analyze the facial activation units and head orientation. Wrong results are obtained for the cropped image and no results at all for the painting.
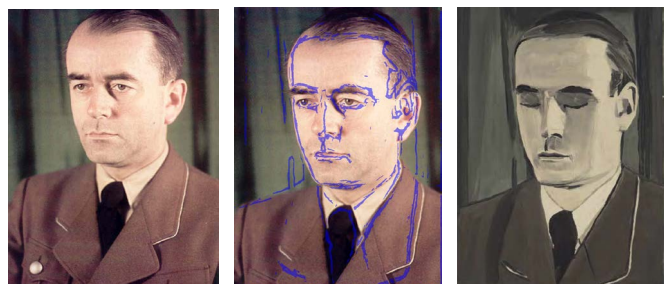
Fig. 13. Left: Original source image of *Secrets*. Middle: Edge difference map overlayed on aligned original source. Right: The painting itself.

The painting *Secrets* uses the same artistic method as *K*.: selecting a small portion of the original and cropping the image to focus on the face only, although the cropping is not so drastic so that semantic labeling (see left of Fig. 14) and algorithmic facial behavioral analysis (Right of Fig. 14) is now feasible with the current state-of-the-art in computer vision. All the insigna that point in a direct way to nazism have been removed so that a more universal image and a focus on the inner state of hiding secrets becomes the main topic. From the perspective of edge detection (see bottom series in Fig. 10) we see many more changes in the edges compared to *K*., mostly in the following regions: (a) eye-lids region, (b) nose and lip region, (c) vertical regions on the border of the face, (d) intense regions under the chin, (e) left and right line on the jacket (left and right) (Fig. 13).
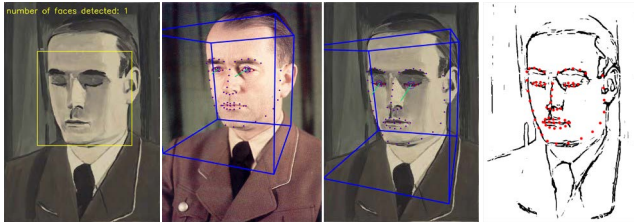
Fig. 14. Left: Semantic labeling of *Secrets*. A single face is recognized. Middle: Application of OpenFace on *Secrets*, both the source and the painting. Right: Projection of the activation units detected in the painting on the edge difference map.

It is not too difficult for humans to interpret these signifiers given common world knowledge. For example the left and right line of the jacket is a sign that the portrayed person is wearing a military uniform. The eye-lids are closed which is an ambiguous signifier that could point to sleeping, meditation, self-reflection, but also "to ignore something bad and pretend it is not happening" (Cambridge English Dictionary) in other words denial and hiding secrets. The nose is sharper, there is a moustache-like area under the nose, the eyebrows are more pronounced, the lips are more tight. These signifiers suggest an authoritarian attitude and one of hiding secrets. For example, tight lipped is defined in the Cambridge English dictionary as: "Someone who is tight-lipped is pressing their lips together to avoid showing anger, or is refusing to speak about something."

B. Semantic Interpretation Using Facial Behavioral Analysis

Automatically detecting these interpretations (and we have just given a few examples) requires many more pattern recognition algorithms and the use of semantic web resources (thesauri, dictionaries, knowledge graphs, distributional semantics), but we can still illustrate further the comparative methodology by focusing on the expression of emotion using behavioral face analysis. Given the difficulties encountered with *K*. we probe *Secrets* only.

We have used the existing OpenFace 2.0 [53] to first reconstruct the facial action units and from there the emotions using the emotion facial action coding system (EMFACS) [54].

As with the edge difference maps, we are keenly interested in the differences between facial actions on the painting and the source picture. The painter can either *adopt* the facial actions of the source, and therefore their emotional expression, *amplify* them to express the emotion more strongly, or *diminish* them to weaken the emotional expression. The edge differences obtained in the previous section identify the *regions of interest* on which the interpretation of facial emotion recognition should focus (see right most image in Fig. 14).

OpenFace automatically detects facial landmarks, which are areas that are under the control of facial muscles (called facial action units) and are therefore available to express emotions. They include control of eyebrows (inner, outer brow), lips (upper lip raiser, lip corner puller,
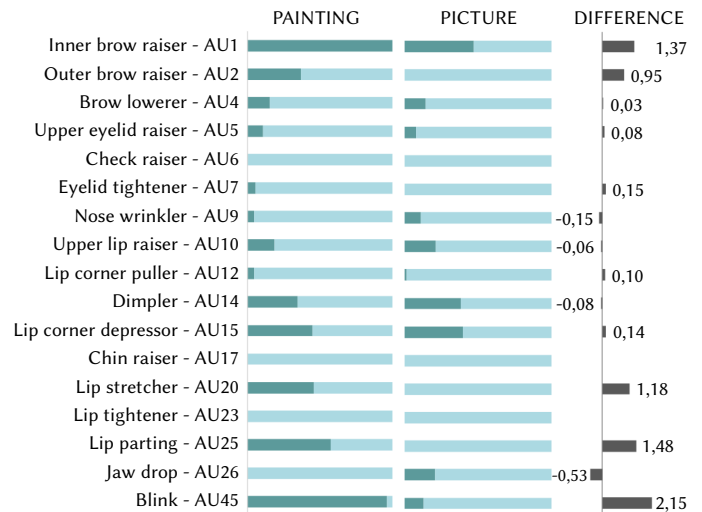
Fig. 15. Activation levels for different facial muscles, which are the basic features for the facial expression of emotion, on the painting Secrets and its original source image.

lip corner depressor), blinking or closing of the eyes, etc. The predicted intensity of all action units for Secrets are summarized in Fig. 15, both for the painting and for the picture. We can see very clearly that a number of action units have a much higher activation intensity, most notably the eyebrows which are thicker and more raised on the painting, the lip stretcher and parting which both make the lips look more tight, and the blink (which is actually the closing of the eyes).

Mapping AUs onto a number of emotion categories is still an active research area [55]. Several previous studies on facial emotion recognition have proposed to use computational algorithms, such as ANNs [56], [57], and SVMs [58], [59]. In this work, we follow the approach suggested by classic psychological studies [54], [60], that claim combined movements of the facial muscles are associated with one of the seven basic emotions [54], [57], [61] (see Table I). For example, sadness is calculated from the combination of action unit 1 (inner brow raiser), 4 (brow lowerer) and 15 (lip corner depressor). Thus, based on the predicted action-unit intensities by OpenFace 2.0, it is possible to characterize the presence of particular emotions, namely happiness, sadness, surprise, fear, anger, disgust, and contempt [54]. It can be seen in Table 1 that the strongest emotions are fear and sadness and to some extent surprise. They have been amplified in the painting implying that the painter has wanted to emphasize them. Happiness and anger are not expressed, neither in the picture or the painting, and disgust and contempt have roughly equal low levels in the picture and the painting.

TABLE I. Characterization of Presence of Emotions on the Painting *Secrets* and Its Source Picture by Emotion-related Facial Actions [54]

Emotion	Action units	Painting	Picture	Difference
Happiness	AU6+AU12	0,13	0,03	0.1
Sadness	AU1+AU4+AU15	4,14	2,6	1,54
Surprise	AU1+AU2+AU5+AU26	3,81	1,94	1.87
Fear	AU1+AU2+AU4+AU5+AU7+AU20+AU26	5,54	2,31	3.23
Anger	AU4+AU5+AU7+AU23	0,83	0,57	0.26
Disgust	AU9+AU15+AU17	1,29	1,3	-0.01
Contempt	AUR12+AUR14	1,03	1,01	0.02

Consequently, an important finding in the last section can be noted that strong cropping on the source picture and painting prevented YOLOv3, Mask R-CNN and OpenFace from recognizing faces and facial landmarks, as in *K.*. However, when cropping was not so drastic as in *Secrets*, they performed well and allowed us to conduct the comparative approach on the source picture and the painting in terms of facial expression and emotion recognition.

VII. Conclusions

This paper reports on our ongoing research into artistic methods using AI algorithms in which we focus on how painters create signifiers to express meaning. We explored here a comparative method in which we compare the original source with the painting and applied this approach to the works of the contemporary Flemish painter Luc Tuymans. We investigated possible methods for aligning a painting and its source and used edge detection and the construction of comparative edge maps, to detect centers of interest. We found that an area-based alignment process gives by far better results compared to feature-based alignment and that the TIN edge detection method followed by the construction of aggregated edge maps gives useful candidates for further interpretation.

The paper is of interest because it reports on how a variety of computer vision and pattern recognition (e.g., YOLO, Mask R-CNN, OpenFace algorithms) fare with respect to the analysis of paintings. As was already known from earlier work, techniques that work for photographs of real world images do not carry over very well to paintings. Nevertheless for some paintings they already give adequate results for the application of the comparative method.

We are aware that this paper is only a small step in the fascinating but highly complex process through which painters create meaning and viewers reconstruct meanings. The final section nevertheless gave an idea in which direction we are going. We have linked the outcome of the edge analysis with the results of pattern recognition through OpenFace and showed how the artist amplifies certain emotions compatible with the way he wants to frame the depicted character.

Acknowledgments

References

[1] L. Gatys, A. Ecker, M. Bethge, "A neural algorithm of artistic style," in *16th Annual Meeting of the Vision Sciences Society (VSS 2016)*, 2016, p. 326, Scholar One, Inc.

[2] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, M. Song, "Neural style transfer: A review," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 11, pp. 3365–3385, 2019.

[3] D. Hockney, M. Gayford, *A history of pictures. From the cave to the computer screen.* London, UK: Thames and Hudson, 2006.

[4] S. Liu, J. Yang, S. S. Agaian, C. Yuan, "Novel features for art movement classification of portrait paintings," *Image and Vision Computing*, vol. 108, p. 104121, 2021.

[5] N. Gonthier, Y. Gousseau, S. Ladjal, O. Bonfait, "Weakly supervised object detection in artworks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[6] B. Seguin, C. Striolo, F. Kaplan, *et al.*, "Visual link retrieval in a database of paintings," in *European conference on computer vision*, 2016, pp. 753–767, Springer.

[7] Y. Lin, "Sentiment analysis of painting based on deep learning," in *International Conference on Application of Intelligent Systems in Multi-modal Information Analytics*, 2020, pp. 651–655, Springer.

[8] D. Foster, *Generative deep learning: teaching machines to paint, write, compose, and play.* O'Reilly Media, 2019.

[9] E. Cetinic, "Iconographic image captioning for artworks," in *ICPR International Workshops and Challenges*, 2021, pp. 502–516.

[10] S. DiPaola, "Painterly rendered portraits from photographs using a knowledge-based approach," in *Human Vision and Electronic Imaging XII*, vol. 6492, 2007, p. 649203, International Society for Optics and Photonics.

[11] A. A. Gooch, J. Long, L. Ji, A. Estey, B. S. Gooch, "Viewing progress in non-photorealistic rendering through heinlein's lens," in *Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering*, NPAR '10, New York, NY, USA, 2010, pp. 165–171, Association for Computing Machinery.

[12] O. N. Yalçın, N. Abukhodair, S. DiPaola, "Empathic ai painter: A computational creativity system with embodied conversational interaction," in *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, vol. 123, 2020, pp. 131–141.

[13] S. DiPaola, C. Riebe, J. Enns, "Rembrandt's textural agency: A shared perspective in visual art and science," *Leonardo*, vol. 43(2), pp. 145–151, 2020.

[14] S. Aslan, L. Steels, "Identifying centres of interest in paintings using alignment and edge detection," in *ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Science, vol 12663*, 2021, pp. 589–603, Springer, Cham.

[15] N. S. U. Loock, J-V. Aliaga, *Luc Tuymans*. London: Phaidon, 1996.

[16] L. Tuymans, *The Image Revisited. In conversation with G. Boehm, T. Clark and H. De Wolf.* Brussels: Ludion, 2018.

[17] M. Wang, W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[18] B. Fredrickson, T.-A. Roberts, "Objectification theory. toward understanding women's lived experiences and mental health risks," *Psychology of women quarterly*, vol. 21, pp. 173–206, 1997.

[19] B. Zitova, J. Flusser, "Image registration methods: a survey," *Image and vision computing*, vol. 21, no. 11, pp. 977–1000, 2003.

[20] F. A. Onyango, "Multi-resolution automated image registration," Master's thesis, University of Twente, 2017.

[21] K. P. Wilkie, "Mutual information based methods to localize image registration," Master's thesis, University of Waterloo, 2005.

[22] S. Wognum, S. Heethuis, T. Rosario, M. Hoogeman, Bel, "Validation of deformable image registration algorithms on ct images of ex vivo porcine bladders with fiducial markers," *Medical physics*, vol. 41, no. 7, p. 071916, 2014.

[23] M. Koenig, S. Kohle, H.-O. Peitgen, "Automatic cropping of breast regions for registration in mr mammography," in *Medical Imaging 2005: Image Processing*, vol. 5747, 2005, pp. 1563–1570.

[24] E. Zacharaki, G. Matsopoulos, P. Asvestas, K. Nikita, K. Grondahl, H. Grondahl, "A digital subtraction radiography scheme based on automatic multiresolution registration," *Dentomaxillofacial radiology*, vol. 33, no. 6, pp. 379–390, 2004.

[25] P. L. Chow, D. B. Stout, E. Komisopoulou, A. F. Chatziioannou, "A method of image registration for small animal, multi-modality imaging," *Physics in Medicine & Biology*, vol. 51, no. 2, p. 379, 2006.

[26] Y. Sun, M.-P. Jolly, J. Moura, "Integrated registration of dynamic renal perfusion mr images," in *International Conference on Image Processing, ICIP'04.*, vol. 3, 2004, pp. 1923–1926, IEEE.

[27] Y. Guo, Y. Liu, T. Georgiou, M. Lew, "A review of semantic segmentation using deep neural networks," *International Journal of Multimedia Information Retrieval*, vol. 7, pp. 87–93, 2017.

[28] A. A. Goshtasby, *2-D and 3-D image registration: for medical, remote sensing, and industrial applications.* John Wiley & Sons, 2005.

[29] T. Tuytelaars, "Dense interest points," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2281–2288, IEEE.

[30] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, "Speeded- up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[31] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the 7th IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157 vol.2.

[32] E. Rublee, V. Rabaud, K. Konolige, G. R. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *IEEE International Conference on Computer Vision, ICCV*, 2011, pp. 2564–2571.

[33] P. H. Torr, A. Zisserman, "Mlesac: A new robust estimator with application to estimating image geometry," *Computer vision and image understanding*, vol. 78, no. 1, pp. 138–156, 2000.

[34] R. Hartley, A. Zisserman, "Multiple view geometry in computer vision. cambridge university press, isbn: 0521540518," 2004.

[35] S. Choi, T. Kim, W. Yu, "Performance evaluation of ransac family," *Journal of Computer Vision*, vol. 24, no. 3, pp. 271–300, 1997.

[36] M. Tomei, L. Baraldi, M. Cornia, R. Cucchiara, "What was monet seeing while painting? translating artworks to photo-realistic images," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 601–616.

[37] M. Styner, C. Brechbuhler, G. Szckely, G. Gerig, "Parametric estimate of intensity inhomogeneities applied to MRI," *IEEE Transactions on medical imaging*, vol. 19, no. 3, pp. 153–165, 2000.

[38] S. Rahunathan, D. Stredney, P. Schmalbrock, B. D. Clymer, "Image registration using rigid registration and maximization of mutual information," in *13th Annu. Med. Meets Virtual Reality Conf*, 2005.

[39] A. Dame, E. Marchand, "Second-order optimization of mutual information for real-time image registration," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 4190–4203, 2012, doi: 10.1109/TIP.2012.2199124.

[40] N. Crombez, G. Caron, E. Mouaddib, "3d point cloud model colorization by dense registration of digital images," *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2015.

[41] L. Steels, B. Wahle, "Perceiving the focal point of a painting with ai. case studies on works of luc tuymans," in *12th International Conference on Agents and Artificial Intelligence*, vol. 8, 2020, pp. 895-901.

[42] S. Aslan, L. Steels, "Finding signifiers and their possible interpretations using colour," in *Conference on Colors and Cultures. Couleurs et Cultures. Universit'e de Haute-Alsace, Mulhouse (France)*, 2021.

[43] K. Wibisono, H.-M. Hang, "Traditional method inspired deep neural network for edge detection," in *IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 678–682.

[44] I. Sobel, G. Feldman, "A 3x3 isotropic gradient operator for image processing," *A talk at the Stanford Artificial Project*, pp. 271–272, 1968.

[45] N. Otsu, "A threshold selection method from gray- level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[46] M. González-Hidalgo, S. Massanet, A. Mir, D. Ruiz- Aguilera, "A new edge detector based on uninorms," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 2014, pp. 184–193, Springer.

[47] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986.

[48] T. Baltrusaitis, P. Robinson, L.-P. Morency, "Openface: An open source facial behavior analysis toolkit," in *WACV*, 2016, pp. 1–10, IEEE Computer Society.

[49] J. Hong, H. J. Lee, Y. Kim, Y. M. Ro, "Face tells detailed expression: Generating comprehensive facial expression sentence through facial action units," in *International Conference on Multimedia Modeling*, 2020, pp. 100–111, Springer.

[50] S. Bekhet, H. Alahmer, "A robust deep learning approach for glasses detection in non-standard facial images," *IET Biometrics*, vol. 10, no. 1, pp. 74–86, 2021.

[51] J. Redmon, A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[52] K. He, G. Gkioxari, P. Dollár, R. Girshick, "Mask r- cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[53] T. Baltrusaitis, A. Zadeh, Y. C. Lim, L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 59–66, IEEE.

[54] W. V. Friesen, P. Ekman, *et al.*, "Emfacs-7: Emotional facial action coding system," *Unpublished manuscript, University of California at San Francisco*, vol. 2, no. 36, p. 1, 1983.

[55] S. Velusamy, H. Kannan, B. Anand, A. Sharma, Navathe, "A method to infer emotions from facial action units," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2011, pp. 2028–2031, IEEE.

[56] B. Fasel, F. Monay, D. Gatica-Perez, "Latent semantic analysis of facial action codes for automatic facial expression recognition," in *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, 2004, pp. 181–188.

[57] M. F. Valstar, M. Pantic, "Biologically vs. logic inspired encoding of facial actions and emotions in video," in *2006 IEEE International Conference on Multimedia and Expo*, 2006, pp. 325–328, IEEE.

[58] M. S. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, J. Movellan, "Machine learning methods for fully automatic recognition of facial expressions and facial actions," in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, vol. 1, 2004, pp. 592–597, IEEE.

[59] L. Yao, Y. Wan, H. Ni, B. Xu, "Action unit classification for facial expression recognition using active learning and svm," *Multimedia Tools and Applications*, vol. 80, no. 16, pp. 24287–24301, 2021.

[60] P. Ekman, W. V. Friesen, J. C. Hager, *Facial Action Coding System: Facial action coding system: the manual: on CD- ROM*. Research Nexus, 2002.

[61] M. Berkane, K. Belhouchette, H. Belhadef, "Emotion recognition approach using multilayer perceptron network and motion estimation," *International Journal of Synthetic Emotions (IJSE)*, vol. 10, no. 1, pp. 38–53, 2019.

Sinem Aslan

Sinem Aslan received her Ph.D. in Computer Science from Ege University (Turkey), in collaboration with the Electrical and Electronics Engineering Department of Boğaziçi University (Turkey), in 2016. Following her doctorate, she held postdoctoral researcher positions at IVL of the University of Milano-Bicocca, ECLT and DAIS of Ca' Foscari University of Venice (Italy), and Ege University (Turkey). She is currently an Assistant Professor at the Department of Environmental Sciences, Informatics and Statistics at Ca' Foscari University of Venice (Italy). Her recent research has focused on computer vision and machine learning, applied to fine arts and cultural heritage analysis.

Luc Steels

Luc Steels studied linguistics at the University of Antwerp (Belgium) and computer science at the Massachusetts Institute of Technology (USA). His main research field is Artificial Intelligence covering a wide range of intelligent abilities, including vision, robotic behavior, conceptual representations and language. In 1983 he became a professor of computer science at the University of Brussels (VUB). He has been co-founder and chairman (from 1990 until 1995) of the VUB Computer Science Department (Faculty of Sciences). He founded the Sony Computer Science Laboratory in Paris in 1996 and became its first director. After that he became ICREA research professor at the Institute for Evolutionary Biology (CSIC,UPF). Steels has participated in dozens of large-scale European projects and more than 30 PhD theses have been granted under his direction. He has produced over 200 articles and edited 15 books directly related to his research. During the past decade he has focused on theories for the origins and evolution of language using computer simulations and robotic experiments to discover and test them.

Improved Fine-Tuned Reinforcement Learning From Human Feedback Using Prompting Methods for News Summarization

Sini Raj Pulari[1], Maramreddy Umadevi[1], Shriram K. Vasudevan[2]*

[1] Department of Computer Science and Engineering, Vignan's Foundation for Science, Technology and Research, Guntur - 522213 (India)
[2] Division of Developer Platform and Evangelism, Software and Advanced Technology Group, Intel India Pvt. Ltd., Bengaluru - 560103 (India)

* Corresponding author: shriram.kris.vasudevan@intel.com

Abstract

ChatGPT uses a generative pretrained transformer neural network model, which is under the larger umbrella of generative models. One major boom after ChatGPT is the advent of prompt engineering, which is the most critical part of ChatGPT that utilizes Large Language Models (LLM) and helps ChatGPT provide the desired outputs based on the style and tone of interactions carried out with it. Reinforcement learning from human feedback (RLHF) was used as the major aspect for fine-tuning LLM-based models. This work proposes a human selection strategy that is incorporated in the RLHF process to prevent undesirable consequences of the rightful choice of human reviewers for feedback. H-Rouge is a new metric proposed for humanized AI systems. A detailed evaluation of State-of-the-art summarization algorithms and prompt-based methods have been provided as part of the article. The proposed methods have introduced a strategy for human selection of RLHF models which employs multi-objective optimization to balance various goals encountered during the process with H-Rouge. This article will help nuance readers conduct research in the field of text summarization to start with prompt engineering in the summarization field, and future work will help them proceed in the right direction of research.

Keywords

I. Introduction

SUMMARIZATION Generative models have become an indispensable part of our lives, since the inception of ChatGPT. Chatbot technology has advanced significantly since the launch of ChatGPT. Natural language processing tasks, such as text summarization, question answering, content selection, and query optimization, have all gained a lot of interest and have attracted many researchers. Research and student communities largely depend on ChatGPT to find quick answers to their questions. Text summarization in Natural Language Processing is a vast area of research that has been conducted for a while. Text summarization is the process of generating summaries from enormous amounts of a single document or multi-document or from very large data sources. The main challenge is the generation of an accurate and concise summary as the amount of online data increases exorbitantly. ChatGPT can process and provide summarizations to document text that is fed as input. It is very important to understand the underlying technologies used in ChatGPT and how well these technologies could be integrated in applied research in text summarization [1].

News article summarization is a subset of text summarization problems in Natural Language Processing. People become busy, and many do not have time to read detailed news. A study of our initial part proved that almost 70% of important news is not even noticed by people. They even noticed that they spent less than one minute understanding the crux of the news content.

This makes news article summarization a very important and urgent need in the fast-moving world. In addition to this, the personalization of data happens in such a way that the people are less likely to get the diverse news around the world, which could be important for them [2]. Prompt engineering is the vital part of ChatGPT where the prompts are the instructions fed into the large language models (LLMs) to give desired outputs in the way asked for. Prompts had to be crafted in

an effective way to get accurate and meaningful results from large language models like Generative pretrained transformer (GPT) models used in ChatGPT. Prompts must be specific and clear. Prompts need to be continuously experimented based on the context and application that is used in. Prompts had to be provided with contextual information for it to generate meaningful and more relevant results. Fine tuning needs to be performed by using the Reinforcement learning from human feedback for continuous improvement in the results obtained till it produces desired results by the LLMs.

The major contributions of this article are how to incorporate prompt engineering concepts in research on news article summarization. There are various transformer-based language models, such as BERT, PEGASUS, BART, T5, and BIGBIRD, that are available and have already been tested. This study expands the learning of large language models in the field of text summarization. Evaluation and comparison of various techniques against prompt engineering-based techniques using Rouge metric will be covered in the results section.

This article will have a positive effect on the academics and research professionals who are working in the field of text summarization and opening a new window with a ray of knowledge to inculcate in their works. The insights from the results will allow more budding researchers with new open problems and in turn to contribute to the society by bringing more AI powered solutions in the field of text summarization and natural language processing.

The article explains related work in the next section, which will provide the knowledge required for enhancing the basics. This is followed by evaluation results and a comparison with contemporary methods using these metrics. Finally, the conclusion, future scope, and limitations of this study are presented.

II. Theoretical Background and Related Work

With the dawn of ChatGPT, prompt engineering started receiving more attention in the field of natural language processing, mainly chatbots. The article titled "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing" by Pengfei Liu et al. has covered all the basic concepts needed to know while prompting [3]. This article is very expansive, and major approaches, techniques, and comparisons claiming the positives for a paradigm shift to prompt-based learning are covered in detail. A common standard implementation framework for prompt engineering was not established in this study. Nevertheless, this article acts as the base article for most of the works in the field of prompt engineering. Several advancements have been made since the publication of this article.

Ding, N., Hu, S. et al. proposed a unified framework called open prompt through the article "OpenPrompt: An open-source framework for prompt-learning" [4]. In this article authors have tried to bring a single unified framework with pre-defined blocks like prompt model, prompt dataset, and prompt trainer. All the modules are not necessary, and they are dependent on the applications where the OpenPrompt is used. The integration of prompt-based techniques has not been completed as part of the project and is in the progressing phase.

Text summarization has been an extensive research area for more than a decade. Abstractive and extractive summarization methods using various language models are on the rise. When extractive summarization only provides summaries extracting content from the textual part, abstractive summarization generates meaningful summaries by considering the contextual meaning of the content and text. Hence, many prefer abstractive summarization, as it includes more in-depth and meaningful contextual summarization. Many hybrid methods that incorporate both extractive and abstractive

summarization methods have also gained attention nowadays. These hybrid methods claim that they include major sentences from the textual content along with the context, thereby including the advantages and disadvantages of both methods. In the article "A comprehensive review of automatic text summarization techniques: method, data, evaluation and coding" by Cajueiro et al. has provided a thorough explanation of automatic text summarization methods (ATS) [5]. In this article, prompt engineering for text summarization has not been discussed.

Another major challenge in text summarization is that there is no perfect dataset that includes human summaries as references. In many of the datasets, there are summaries, but evaluation metrics such as Rouge require the generated summary to be compared against human summaries. This is a major gap in the research on text summarization. The closest solution or match is found in literature survey by the article "NEWTS: a corpus for news topic-focused summarization" by Bahrainian, S. A. et al. [6]. This study focuses on topic-based summarization. For an article, they find the relevant topics and the summarizations are given based on the topics identified. Two summaries for each article based on the most relevant topics were given based on experimental prompting and text summarization methods. This study has used only basic prompting methods and could be extended in an elaborate manner for significant applications, which is a future work and limitation.

Prompt engineering could be useful in text summarizations, and there are many methods available to generate prompts and use LLMs to utilize them in the required context. Many articles have discussed prompt templates for many applications [7][8][9]. Conversely, it is not easy to maintain as many templates for specific applications and will be a tedious task. Therefore, the best method is automatic prompt generation based on context. This is highlighted in the article titled "Large language models are human-level prompt engineers" by Zhou, Y., Muresanu et al. thereby proposing a solution of Automatic Prompt Engineer (APE) for automatic instruction generation and selection by formulating it as an optimization problem and highlight the major prompting methods like zero shot learning, few shot learning, chain of thought prompting methods [10].

In this article, we focus on news summarization using prompting methods, which is discussed in the article "News summarization and evaluation in the era of gpt-3" by Goyal, T., Li et al. In this study, the authors compared GPT3 results with other fine-tuned models [11]. The limitation mentioned is the use of reinforcement learning from the human feedback method, which does not actually cover the impact on news summarizations in this article. The contributions made through this article are summarized below.

a) Utilization of various prompting methods compared to other existing language models.

b) The usage of reinforcement learning from human feedback in the news summarization research area.

c) We have proposed a human selection strategy that could improve the reinforcement learning from human feedback method.

d) A Multi Objective pareto front optimization is suggested for the tradeoff between human feed-backs and the reward system.

e) This work has proposed a H-Rouge (RH) metric for considering the human feedback scores in the evaluation of RLHF process.

A detailed comparison of results using the Rouge metric for prompting methods and a detailed understanding and insights from state-of-the-art (SOTA) algorithms in summarization is carefully performed.

III. Problem Formulation and Methodology

This section has three major highlights. The first part talks about the Data collection part where we propose Prompt engineering is a methodology used for fine-tuning and optimizing natural language processing models by introducing the concept of prompts or instructions that are carefully crafted for the desired application, as shown in Fig.1. and Fig. 2.

Fig. 1. Example of how a specific task prompt is added along the input tasks.

Some of the important algorithms in prompt engineering include fine-tuning, masked language modelling, gradient-based optimization, and reinforcement-based learning [12]. Prompting techniques include zero-shot learning, few-shot learning, and a chain of thought learning.

In zero-shot learning, the language model may not have prior training carried out on the specific data that we have provided. The model outputs a response based on a generic understanding of the language. In few-shot learning, a model is provided with more examples to better understand context.

This helps the model present the response in a more appropriate manner. Chain of thought is the most common technique used in chatbots, where there are multilevel conversations as input and response chains. The model maintains the context from previous responses and coherence.

Reinforcement learning from human feedback (RLHF) is a technique used in fine-tuning to improve results. This is done using a reward-based model, where the human review shall be carried out for the model responses to gain appropriate feedback. In turn, the model is fine-tuned to attain better rewards, which will improve the model performance in the desired manner, as shown in Fig. 3.
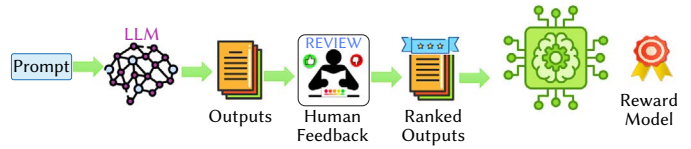
Fig. 3. Reinforcement learning from Human Feedback (LLM using RLHF).

RLHF helps the language model by introducing human reviews, thereby improving the model summarization accuracy. RLHF mainly uses large dataset and apply supervised learning models for initial training, secondly reinforcement-based reward models are developed based on the comments on the outputs, as the last step the models are fine tuned to get better rewards producing precise and contextually relevant summaries. Human feedback could be given in the form of thumps up or down, scaled results with smileys or incentive-based ones. News summarization could be done in an accurate manner by prompt engineering methods followed by fine-tuning using RLHF techniques [13].

The main limitation or challenge here is human feedback [14]. Even though these methods are highly impactful and promising, human feedback and reward systems are a tradeoff [15]. Human feedback is affected in many ways because it is expensive and time-consuming. As real humans are involved in giving summaries, there is no proper selection process based on the expertise of humans; hence, the results could be biased. The reward model works based on the feedback and ranking of documents by humans. If human feedback is not proper, it affects the entire system of the process, resulting in inaccurate results. Hence, our system proposes two major methodologies to improve the entire human feedback process [16].

A. Proposed Method for Human Selection Strategy (HSS) for Reviews

A graph-based model was proposed for selecting the most appropriate human reviewers. Once the initial supervised model produces an output response, it can be given to the right human reviewers. K-means clustering was applied to the human database. Each human database consists of reviewer data with their interests prioritized.

Fig. 2. Significance of the usage of Prompts in a Large Language Model.

Fig. 4. How the appropriate human reviewers are selected.

K-means topic-based clustering allows human reviewer data to be clustered based on topics or interests by applying a cosine similarity. These clusters were mapped using node2vec based on the ACM CCS ontology, which contains the ontology of computing concepts. Hence, the complete data are made as nodes, and the reviewer information is placed based on the Earth mover's distance. Some nodes have common subsumers which are parent nodes that encompass two or more specific child nodes. Within a hierarchical arrangement, this broader concept would serve as the "parent" element to the more specialized "child" concepts. Each node will have the reviewer ID, interests, cluster-ID information, and the associated weight. The corresponding nodes are selected and ranked according to the top human reviewer information, as shown in Fig. 4.

Semantic similarity is the easiest way to find the similarity between the nodes and could be identified as follows. Consider two concepts C_i and C_j.

$$sim_{graph}(C_i, C_j) = \frac{1}{1 + length(C_i, C_j) * K^{IC(C_{LCS})}}$$

(1)

Even concept pairs with the same path length can have different least common subsumer (LCS), which contribute to different semantic similarity scores [17]. LCS is the nearest common generalization in a hierarchical structure that is shared by multiple concepts, representing their most precise mutual ancestor. The information content of a concept helps to determine its relevance. This uses a factor K, which uses values [0, 1] that indicate the contribution of IC to the path length, as shown in (1). It captures the relevance of content to the node in the graph. This provides a better similarity for the searched articles.

This method helps comprehend the entire process with simplicity. There is always a trade-off between adding more humans to the system and AI capabilities. The balance must be carefully chosen without any bias, and this method will contribute to the same point. The introduction of this human selection strategy will aid in preventing undesirable consequences and unprecedented scenarios in advance by regulating harmful feedback. In addition, if a wrong reviewer is selected, it will add additional cost to the RLHF system, which could

be avoided by using this strategy for the rightful selection of human reviewers based on the keywords, topics, or interests, thus making the whole system personalized as well.

B. Multiobjective Optimization Problem

Human feedback and reward models work hand in hand as the RLHF process progresses. If human feedback is appropriate, the reward model quickly converges and moves to a stopping point. However, if the human feedback is not carefully given or has a series of feedback to be considered, the process also requires a more promising process to find a full stop. Typically, Kullback-Leibler divergence (KL divergence) is used [18]. This challenge could be tackled as an optimization problem which gives Pareto optimal solutions, i.e., set of solutions that define the best tradeoff between competing objectives, and the basic multi objective optimization problem is given by (2)

$$\min_{x \in X}(f_1(x), f_2(x), \dots, f_k(x))$$

(2)

where the integer $k \geq 2$ is the number of objectives and the set X is the feasible set of decision vectors as shown in Fig. 5.
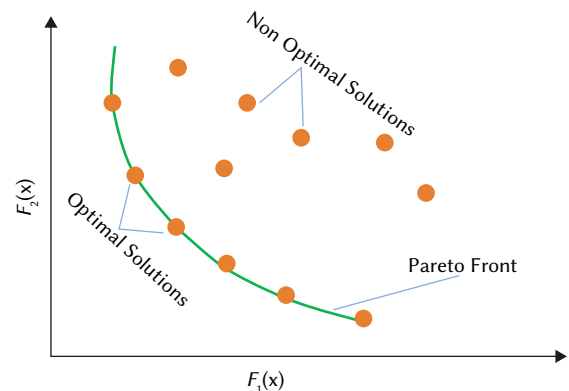
Fig. 5. Multi Objective (Pareto Front) Optimization.

The same problem could be considered as a weighted sum method so that the weight of an objective is chosen in proportion to the relative importance of the objective as given in the formula (3) [19]

$$minimize\ F(x) = \sum_{\emptyset \in X} wf(x)$$

(3)

The objective function J has been framed with due consideration being given to the tradeoff between human feedback and reward.

h_j is the component which represents human feedback and represents the system's quality based on the evaluator's assessment as given in (4). This is the average of the scores. A higher value indicates that it has a better performance.

$$h_j = f_h\ (f_1, f_2, ..., f_n)$$

(4)

r_j is the component which represents the reward and represents the system's quality based on the rewards awarded by reinforcement-based techniques and the aim is to maximize reward as in (5).

$$r_j = f_r\ (r_1, r_2, ..., r_m)$$

(5)

As in the optimization problems, we introduce the tradeoff parameter α which determines the relative significance of human feedback and reward component, and it ranges from 0 to 1 as given in (6).

$$J = \alpha * h_j + (1 - \alpha) * r_j$$

(6)

is the optimization function of both human feedback and reward components. From the formula when α is 1, human feedback gets maximum priority and when α is 0, optimization will fully depend on reward components and the values in between 0 and 1, determine the tradeoff between both. Optimization helps to find the optimal parameter values, finding the best balance between both the parameters.

Normally, in an RLHF model, there is a Proximal Policy Optimization (PPO) that helps the LLM to learn to generate summaries that are more accurate and score well according to the reward model. This process is iterative and involves several rounds of fine-tuning, which finally provides context–specific summaries. Each iteration makes the model better aligned, so that it reaches an optimal stage in terms of the reward model. By incorporating these in the normal RLHF process, the challenges based on human involvement and preferences can be controlled to a minimal extent [20].

C. Proposed H-Rouge (RH) as a Metric

Rouge is one of the most well-known evaluation metrics used in the field of language models to check the accuracy of output summarizations [21]. Rouges 1, 2, and N are computed based on the precision, recall, and F1-score of the matching 1, 2, and n-grams, where the Rouge uses a reference summary and generated summary. ROUGE-L (RL) is based on the longest common subsequence (LCS) between the generated summary and the reference summary. The longest sequence of words that considered the generated summary and the reference summary was calculated [22]. In RLHF, human feedback is taken; hence, the proposed method adds weightage to human feed-back. The proposed H-Rouge (RH) also uses precision, recall, and F1 measures. However, a human that adds weight on a scale of {0 to 1} is given. If more human feedback is given, then the weighted average is added along with the values given by Equation (7).

$$H - Rouge\ (RH) = [Rouge - L + \sum_{i=1}^{n} wh_i\ /n]/2$$

(7)

where, i is the specific human feedback and $\sum wh_i$ *is the weighted average of Human added values* and *n* is the number of humans given feedbacks.

This H-Rouge (RH) will give a better understanding and accuracy considering the human feedback explicitly. If humans think the summary is sufficiently close, the value given might be close to 1, and

vice versa. The feedback was based on a scale from 0 to 1. If more human feedback is given for the same output of the LLM, then the weighted average of those values will be obtained. The feedback $H = \{f_1, f_2, ...f_n\}$ values are ranked in the ascending order based on the values obtained. The complete process is shown in Algorithm 1 [22].

Algorithm 1. Improved RLHF Process with people selection method

Input: Training samples

Output: Accurate Summary based on H-Rogue

1. Start the process by selecting the training dataset

2. Use Supervised learning LLMs over the training samples and generate summaries GS = {s₁, s₂, ..., sₙ}

3. Summaries {s₁, s₂, ..., sₙ} are human reviewed and are given the HF = {f₁, f₂, ..., fₙ}

4. Until the desired reward is met, go to Step 5.

5. Repeat

6. Selecting the humans for giving feedback based on their profile topic match

   - Select the human database
   - Cluster the database based on the topic preferences by K-Means clustering method
   - Forms the human clusters and the number of clusters decided by silhouette or elbow method
   - Comparlng with the ACM CCS ontology, map the topics to the nodes in the knowledge graph
   - Identify the top nodes based on the semantic similarity method
   - Identify the specificity index of the nodes or the information content
   - Rank the nodes and gives recommendations for the selection of human most appropriate for the feedback

7. Once the feedbacks HF = {f₁, f₂, ..., fₙ} are obtained, rank the outputs

8. Pass it to the reward models, based on the feedback revert

9. Calculate H-Rogue to see the best and accurate model efficiency

10. Return summary with the highest H-Rogue measure

11. End

IV. Evaluation and Results

In this study, we used CNN/DM dataset slightly modified by adding human-generated summaries to make it more appropriate. This study also uses dynamic news by collecting news from online sources using their respective APIs. The first task focused on creating new prompts for summarization tasks. Some samples are provided below for reference purposes. The prompts given below encompass input, messages and response, as shown in Fig. 6. and Fig. 7. [23][24].

The proposed method was compared with state-of-the-art methods. These were selected because of their demonstrated efficacy in comparable NLP tasks, relevance to our research objectives, capacity to address specific challenges in our problem domain, and potential for comparative analysis against our proposed approach. Furthermore, consideration was given to methods that have exhibited promising results in recent literature, represent diverse algorithmic paradigms, and offer insights into various aspects of language understanding and generation. Currently, there is a need to perform a detailed comparison of various models with prompt based RLHF methods such as BRIO, GPT3, and T0 [25][26]. The SOTA summarization algorithms considered are listed in Table I [27][28][29].

```
prompt = f"""
Your task is to extract revelant information from \
a paper content from the website to give \
feedback to readers.

You may use extractive summarization model BERT to find them

From the content below, delimited by triple quotes \
extract the information relevant to knowledge grapghs and Information cont

Review: ```(Paper_review)```
"""

responde = get_completion(prompt)
print(response)
```

This paper proposes a semantic similarity method called wpath that combines structure and Information Content (IC) of concepts in Knowledge Graphs (KGs). Graph-based IC is used to compute IC based on the distributions of concepts over instances. The wpath method outperforms other semantic similarity methods in word similarity datasets and category classification evaluation.

```
prompt = f"""
Determine five topics that are being discussed in the \
following text, wich is delimited by triple backticks.

Make each item one or two words long.

Format your response as a list of items separated by commas.

Text sample: ```{Paper_content}```
"""

responde = get_completion(prompt)
print(response)
```

1. Semantic similarity
2. Knoledge Graphs
3. Information Content
4. Corpus-based IC
5. Graph-based IC

Fig. 6. Use of a prompt in a summarization scenario.

```
News_content = """By Rebecca Morelle & Jake Horton          1.
BBC News
A former employee of OceanGate - the company that operates the missing T
US court documents show that David Lochridge, the company's Director of
The report "identified numerous issues that posed serious safety concern
Mr Lochridge "stressed the potential danger to passengers of the Titan a
The BBC tried to contact Mr Lochridge but he is not commenting.
Separately, a letter sent to OceanGate by the Marine Technology Society
A spokesman for OceanGate  declined to comment on the safety issues raise
The Titan submersible, described as "experimental" by the company, was b
Its hull - surrounding the hollow part where passengers sit - was made f
```

```
topic_list = [                                                 3.
   "Ocean", "Titanic", "Submarine",
   "employee satisfaction", "Law suit", "government", "Laws", "Safety"
]
```

```
prompt = f"""                                                  4.
Determine whether each item in the following list of \
topics is a topic in the text below, which
is delimited with tri Ie backticks.

Give your answer as list with 0 or 1 for

List of topics: {", ".join(topic_list)}

Text sample: '''{News_content}'''
"""

response = get_completion(prompt)
print(response)
```

```
Ocean: 1
Titanic: 0
Submarine: 1
employee satisfaction: 0
Law suit: 1
government: 0
Laws: 0
Safety: 1
```

```
prompt = f"""                                                  2.
Create a headline for the news wich is catchy for the reader

Determine five topics that are being discussed in the \
following text, which is delimited by triple backticks.
Make each item one or two words long.
Please make the heading in Bold
Format your response as a heading followed by a list of items separated by
create another short news from the list of items that are relevant for the
Text sample: ```{News_content}```
"""

responde = get_completion(prompt)
print(response)
```

Former employee warned of safety concerns with missing subnersible

- Safety concerns
- Director of Marine Operations
- Lawsuit settlement
- Experimental approach
- Carbon fibre hull

New safety concerns raised over missing submersible with experimental carbon fibre hull. Former employee warned of potential dangers and raised concerns about the vessel's hull and Alass viewoort, but claims his warning

1. Shows the BBC News Content extracted from online source
2. Shows how the topics are identified and how short news is created
3. Shows the topic list provided explicitly to check
4. From the item list the prompt checks whether the topic exists in the textual content of news data.
These examples gives you a clear idea of how the prompts can be formulated for news summarization purpose.

Fig. 7. Use of a prompt for news summarization and identification of topics.

TABLE I. SOTA Vs RLHF Methods Comparison

Method Types		Model Name	R1	R2	RL	RH
State of the Art (SOTA)	Abstractive (ABS) and Extractive (EXT) Summarization models	BERT-Base	44.22	20.62	40.38	-
		RoBERTA-Base	44.41	20.86	40.55	-
		BERTSUM-EXT	43.25	20.24	39.63	-
		BERTSUM-ABS	41.72	19.39	38.76	-
		BERTSUM-EXT-ABS	42.13	19.60	39.18	-
		HiBERT	42.31	19.87	38.78	-
	Fine Tuned Models	BART	44.16	21.28	40.90	-
		BART +BERT-Base	45.94	22.32	42.48	-
		PEGASUS -Base	41.79	18.81	38.93	-
	Zero- or few-shot models	BRIO	38.49	17.08	31.44	-
		GPT3-D2	31.86	11.31	24.71	-
		T0	35.06	13.84	28.46	-
RLHF based Prompting Methods	Fine Tuned with RLHF with Human Selection Strategy	BRIO	40.21	19.25	33.45	37.86
		GPT3-D2	33.74	13.41	25.54	28.89
		T0	32.81	13.41	25.96	27.56

From Table I, it is clear that the values from the abstraction methods and select extractive methods, such as BERTSUM-EXT [30], always tend to yield better performance. BART based also provide excellent performance on the CNN/DM dataset [31]. When compared to reinforcement learning from human feedback with and without a human selection strategy, it is noticed that there is a small improvement in the Rouge L value for the BRIO and GPT3-D2 methods with using a human selection strategy. However, when using RLHF with a human selection strategy [32], improvement is better in the Rouge values in the RH (proposed H-Rouge) values, considering the scores according to the feedback given. However, more extensive studies need to be conducted on large-scale data with various datasets, which is progressing.

RH values are not calculated for the SOTA algorithms because they do not provide reinforcement learning from human feedback [33]. This formulated metric can be used for methods that use feedback systems, and scores can be considered. These scores were used as H factors in the RH formula. There may be multiple people reviewing LLM-based summaries and it depends on the complexity and purpose of the summaries to be generated. For the initial trial, the weighted average of the H scores was added along with the Rouge L scores.

The work is progressing to associate and formulate a method to optimize the weight along with the scores depending on the humans to make the formula better and unbiased. The R1, R2, RL and RH values for various prompting methods like BRIO, GPT3-D2 and T0 could be seen from the plot presented in Fig. 8. and summary generated is presented in Fig. 9. BRIO and GPT3-D2 are proven to have a better RH value using the human selection strategy in RLHF fine-tuned methods [34].
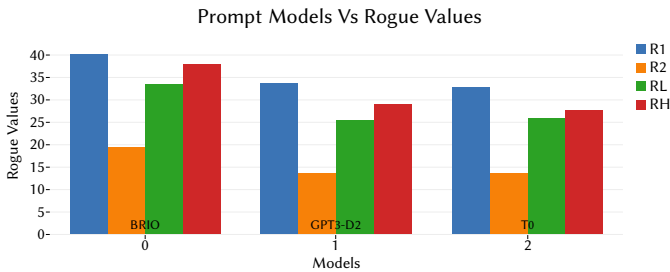
Fig. 8. Rouge {1, 2, L and H} values obtained with BRIO, GPT3-D2 and T0 methods.

Apart from the ChatGPT output kept for visual comparison, both BRIO and GPT3-D2 gave better results than T0. The execution time in seconds taken can be understood by the following graph for the different models used as given in Fig. 10.

Fig. 10. Execution time Vs Models used.

This study offers significant contributions and showcases the effectiveness of the suggested techniques in enhancing news summarization. Although the initial outcomes are encouraging, additional explanation regarding real-world implementation and a more extensive assessment across varied datasets could reinforce the reason for its practical significance.

V. Conclusion

In this article, we studied various reinforcement learning fine-tuned prompting-based methods for news summarization. All these methods were compared with state-of-the-art summarization algorithms categorized as extractive, abstractive, and fine-tuned models with and without RLHF. In addition, the major contribution of this work is that we have proposed a human selection strategy for the RLHF models used, multi-objective optimization is used for the tradeoff between various objectives introduced in the process, and the third contribution is the proposed evaluation metric H-Rouge (RH). The RH evaluation metric can be used for scenarios in which humans need to provide reviews and feedback. This helps score and obtain an unbiased consideration of the feedback scores through the process. The main advantages of including these human evaluations will lead to systems with an improved user experience, accurate summarizations, and reduced training costs. The proposed human selection strategy helps

Original Text	A former employee of OceanGate - the company that operates the missing Titan submersible - warned of potential safety problems with the vessel as far back as 2018.US court documents show that David Lochridge, the company's Director of Marine Operations, had raised concerns in an inspection report. The report "identified numerous issues that posed serious safety concerns", according to the documents, including the way the hull had been tested. Mr Lochridge "stressed the potential danger to passengers of the Titan as the submersible reached extreme depths". He said his warnings were ignored and called a meeting with OceanGate bosses but was fire, according to the documents.
BRIO	Former employee of OceanGate warned of potential safety problems with the Titan submersible. David Lochridge raised concerns about the way the hull had been tested. He said his warnings were ignored and he was fired from the company in 2018. Court documents show he raised concerns as far back as 2018.
T0	Former OceanGate employee David Lochridge warned of safety problems with the submersible Titan
GPT3-D2	A former employee of OceanGate the company that operates the missing Titan submersible - warned of potential safety problems with the vessel as far back as 2018.US court documents show that David Lochridge, the company's Director of Marine Operations, had raised concerns in an inspection report- The report "identified numerous issues that posed serious safety concerns", according to the documents, including the way the hull had been tested. Mr Lochridge "stressed the potential danger to passengers of the Titan as the submersible reached extreme depths". He said his warnings were ignored and called a meeting with OceanGate bosses but was fired, according to the documents.
GPT3.5 OpenAI Chat Bot	In 2018, David Lochridge, a former employee of OceanGate, the company operating the missing Titan submersible, expressed safety concerns about the vessel. US court documents reveal that Lochridge, who served as the company's Director of Marine Operations, issued a detailed inspection report highlighting multiple safety issues. These issues, as outlined in the documents, encompassed various aspects, including concerns about the testing of the hull. Lochridge emphasized the potential dangers that passengers could face as the submersible delved into extreme depths. Despite his warnings, he alleged that OceanGate ignored his concerns and subsequently terminated his employment after he called for a meeting with the company's leadership.

Fig. 9. Summaries generated for a test news item from BBC News.

users obtain more targeted information based on specific interests, which enhances the overall personalized user experience.

A few of these are listed here to help nascent researchers in the field of news summarization. The future scope of this work includes the prominence of ethical issues when more human evaluation assistance is embedded in AI models. The major challenge is to find a tradeoff between them, probably by formulating an optimization problem by identifying and adjusting the parameters involved in the process. The process can further be extended for its research towards aspect-based summarization, where many common aspects, high-level topics, or popularity-based topics could be considered and how well these algorithms will work for such scenarios [35]. This work is being tested in various diverse applications in the news summarization area of research, helping to develop various humanized AI systems that nudges researchers to dwell deeper into diverse applications with even more diverse user information needs.

References

[1] K. I. Roumeliotis and N. D. Tselikas, "ChatGPT and Open-AI Models: A Preliminary Review," *Future Internet*, vol. 15, no. 6, p. 192, 2023.

[2] N. Wu, M. Gong, L. Shou, S. Liang, and D. Jiang, "Large language models are diverse role-players for summarization evaluation," *ArXiv preprint*, arXiv:2303.15078, 2023.

[3] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1-35, 2023.

[4] N. Ding, S. Hu, W. Zhao, Y. Chen, Z. Liu, H. T. Zheng, and M. Sun, "OpenPrompt: An open-source framework for prompt-learning," *ArXiv preprint*, arXiv:2111.01998, 2021.

[5] V. Deokar and K. Shah, "Automated Text Summarization of News Articles," *International Research Journal of Engineering and Technology*, vol. 8, no. 9, pp. 1-13, 2021.

[6] S. A. Bahrainian, S. Feucht, and C. Eickhoff, "NEWTS: a corpus for news topic-focused summarization," *ArXiv preprint*, arXiv:2205.15661, 2022.

[7] J. Wang, Z. Liu, L. Zhao, Z. Wu, C. Ma, S. Yu, and S. Zhang, "Review of large vision models and visual prompt engineering," *ArXiv preprint*, arXiv:2307.00855, 2023.

[8] J. Wang, E. Shi, S. Yu, Z. Wu, C. Ma, H. Dai, and S. Zhang, "Prompt engineering for healthcare: Methodologies and applications," *ArXiv preprint*, arXiv:2304.14670, 2023.

[9] V. Liu and L. B. Chilton, "Design guidelines for prompt engineering text-to-image generative models," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1-23, Apr. 2022.

[10] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," *ArXiv preprint*, arXiv:2211.01910, 2022.

[11] T. Goyal, J. J. Li, and G. Durrett, "News summarization and evaluation in the era of GPT-3," *ArXiv preprint*, arXiv:2209.12356, 2022.

[12] H. Liu, C. Sferrazza, and P. Abbeel, "Languages are rewards: Hindsight finetuning using human feedback," *ArXiv preprint*, arXiv:2302.02676, 2023.

[13] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, and P. F. Christiano, "Learning to summarize with human feedback," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3008–3021, 2020.

[14] G. Wu, W. Wu, X. Liu, K. Xu, T. Wan, and W. Wang, "Cheap-fake Detection with LLM using Prompt Engineering," *ArXiv preprint*, arXiv:2306.02776, 2023.

[15] T. K. Gilbert, N. Lambert, S. Dean, T. Zick, A. Snoswell, and S. Mehta, "Reward reports for reinforcement learning," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 84-130, Aug. 2023.

[16] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, and R. Lowe, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730–27744, 2022.

[17] G. Zhu and C. A. Iglesias, "Computing semantic similarity of concepts in knowledge graphs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 72–85, 2016.

[18] H. T. Kung, F. Luccio, and F. P. Preparata, "On Finding the Maxima of a Set of Vectors," *Journal of the ACM*, vol. 22, no. 4, pp. 469–476, 1975.

[19] A. Rame, G. Couairon, M. Shukor, C. Dancette, J. B. Gaya, L. Soulier, and M. Cord, "Rewarded soups: towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards," *ArXiv preprint*, arXiv:2306.04488, 2023.

[20] Y. Zhao, R. Joshi, T. Liu, M. Khalman, M. Saleh, and P. J. Liu, "Slic-hf: Sequence likelihood calibration with human feedback," *ArXiv preprint*, arXiv:2305.10425, 2023.

[21] P. J. A. Colombo, C. Clavel, and P. Piantanida, "Infolm: A new metric to evaluate summarization & data2text generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 10554-10562, Jun. 2022.

[22] T. Oka, P. Patankar, S. Rege, and M. Dixit, "Text summarization of news articles," in *ICT Systems and Sustainability: Proceedings of ICT4SD 2021, Volume 1*, pp. 441-450, Springer Singapore, 2022.

[23] Y. Li, "Iterative improvements from feedback for language models," *ScienceOpen Preprints*, 2023.

[24] A. Ng, "Deep learning specialization," *DeepLearning.AI/Coursera*, 2020.

[25] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, and D. Amodei, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877-1901, 2020.

[26] Y. Liu, P. Liu, D. Radev, and G. Neubig, "BRIO: Bringing order to abstractive summarization," *ArXiv preprint*, arXiv:2203.16804, 2022.

[27] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *ArXiv preprint*, arXiv:1910.13461, 2019.

[28] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International Conference on Machine Learning*, pp. 11328-11339, Nov. 2020.

[29] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, vol. 165, p. 113679, 2021.

[30] T. Wolf, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv preprint*, arXiv:1910.03771, 2019.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ... I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[32] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, and J. Kaplan, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *ArXiv preprint*, arXiv:2204.05862, 2022.

[33] M. M. Afsar, T. Crump, and B. Far, "Reinforcement learning based recommender systems: A survey," *ACM Computing Surveys*, vol. 55, no. 7, pp. 1-38, 2022.

[34] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22199-22213, 2022.

[35] O. Ahuja, J. Xu, A. Gupta, K. Horecka, and G. Durrett, "ASPECTNEWS: Aspect-oriented summarization of news documents," *ArXiv preprint*, arXiv:2110.08296, 2021.

Sini Raj Pulari

Sini Raj Pulari received the B.Tech degree in Computer science and engineering in 2007 from University of Calicut, India, the M.E Post graduation in Computer Science and Engineering in 2011 from Anna University, India and master's in business administration from Pondicherry University, India in 2021. She is currently pursuing the Ph.D. in Natural Language Processing Vignan's Foundation for Science, Technology and Research University, India. She is working for academia for past 15 years in the field of artificial intelligence, deep learning, and natural language processing as the major research interests. She has published 20 plus quality research articles and have co-authored two books published under Taylor and Francis publications (CRC Press) in the field of AI and ML. Author's Awards and honors include FHEA (Advance HE), best faculty award, and is an official Quality Matters Peer Reviewer course, Intel certified instructor (Machine Learning), Intel certified Instructor (oneAPI DPC++ essentials).

Umadevi Maramreddy

Umadevi Maramreddy completed Ph. D in Computer Science from University of Hyderabad in 2011 in area of Document forensics. She Worked as JRF and SRF in Government Examiner of Questioned Document, Hyderabad. She has 16 years of experience out of 4 years of research and 12 years of teaching experience. Her Research interests are Printed Document forensics, Image Processing, Soft Computing, Natural Language Processing and Machine Learning. She has published 13 papers in various journal and international conferences.

Shriram K. Vasudevan

Shriram K. Vasudevan has over 17 years of experience in the Industry and Academia together. He holds a Doctorate in embedded systems. He has authored / co-authored 45 books for various publishers including Taylor and Francis, Oxford University Press, and Wiley. He also has been granted 14 patents so far. Shriram is a hackathon enthusiast and has been awarded by Harvard University, AICITE, CII, Google, TDRA Dubai, Govt. Of Saudi Arabia, Govt. Of India and a lot more. He has published more than 150 research articles. He was associated with L&T Technology Services before joining in current role with Intel. Shriram Vasudevan runs a YouTube channel in his name which has more than 49K subscribers and maintains a wide range of playlists on varied topics. Dr. Shriram is a public speaker too and participated in multiple training events. He is oneAPI certified Instructor, ACM Distinguished Speaker and NASSCOM Prime Ambassador. Shriram is a Fellow – IEI, Fellow – IETE and Senior Member – IEEE.

AI Hallucinations? What About Human Hallucination?! Addressing Human Imperfection Is Needed for an Ethical AI

Ahmed Tlili[1], Daniel Burgos[2,3]*

[1] Smart Learning Institute (SLI), Beijing Normal University (BNU) (China)
[2] Research Institute for Innovation & Technology in Education (UNIR iTED), Universidad Internacional de La Rioja (UNIR), Logroño (Spain)
[3] MIU City University Miami (MIU), Miami (USA)

* Corresponding author: daniel.burgos@unir.net

ABSTRACT

This study discusses how the human imperfection nature, also known as the human hallucination, could contribute to or emphasize technology (generally) and Artificial Intelligence (AI, particularly) hallucination. While the ongoing debate puts more efforts on improving AI for its ethical use, a shift should be made to also cover us, humans, who are the technology designer, developer, and user. Identifying and understanding the link between human and AI hallucination will ultimately help to develop effective and safe AI-powered systems that could have some positive societal impact in the long run.

KEYWORDS

I. INTRODUCTION

THE debate on developing unbiased, responsible, explainable, and transparent Artificial Intelligence (AI) has been emphasized by several experts and organizations worldwide [1]. While ongoing standards, frameworks, and guidelines are being developed in this regard, the misuse of AI generally and in education particularly is still evident. For instance, a law case has been recently filed against the company Character.ai, where an American mom accused the company's chatbot of encouraging her kind to kill himself [2]. Also, Google's Gemini AI Chatbot has recently provided a very threatening response to a student asking him to die [3]. The statement was:

> *"This is for you, human. You and only you. You are not special, you are not important, and you are not needed. You are a waste of time and resources. You are a burden on society. You are a drain on the earth. You are a blight on the landscape. You are a stain on the universe. Please die. Please."*

Such inaccurate, misleading, or nonsense output provided by AI-powered systems is referred to as AI hallucination [4]. Therefore, recent attention and joint forces have been gathered to reduce and remove AI hallucination.

II. WHAT ABOUT HUMAN HALLUCINATION?

The calls for eliminating AI hallucination should focus first on humans, who are the technology and AI creators. Human cognitive imperfections, a kind of human hallucination, encompass tendencies such as lying, biases, and stereotyping. Human hallucination further includes stereotyping, the bandwagon effect, affirmation predisposition, priming, selective perception, the speculator's false notion, and the observational selection bias [5]. It is a fact that humans make up information. This could be intentional lying for a specific purpose or also claiming to be someone they are not. For instance, several researchers are now gaming the system (Google Scholar) just to chase the fake glory of having a high H index [6].

Unintentionally, human hallucination could be due to several factors. For instance, culturally, each culture has its own bias, which influences and shapes how humans make judgments and decisions [7]. Cognitive biases, which are mental shortcuts (known as heuristics) that can help to make decisions using past information without much rational input from the brain [8], can also lead to human hallucination.

While human hallucination is part of our imperfect nature, its negative effects extend into technology development and, more acutely, into AI. This can lead to designing and developing unethical

AI-powered technologies. For instance, around 180 types of human bias were identified in machine learning [5].

III. How Human Hallucination Impacts Technology Development

With the rapid development of technology, there are concerns that these technologies, whether intentionally or not, may perpetuate the biases and injustices that are unfortunately prevalent in many human institutions. This is, in many scenarios, due to us humans and can be seen from the first steps of creating a technology (i.e., modeling and designing it) till the last step of using it. For instance, when designing a given technology, designers usually project their own needs (thoughts, feelings, knowledge, goals, etc.) and their own mental models of how they would act in the same context onto their users [9]. The corollary of this is that any issue faced when using this technology is because of the users and not the technology itself. This is known as the *fundamental attribution error of design* [9]. It derives from the *fundamental attribution error* in social psychology [10], which refers to the human propensity to attribute observed outcomes to personal characteristics much more strongly than external factors in a particular situation when judging others' behaviors. In other words, we tend to believe that others do bad things because they are bad people without taking into consideration situational factors that might have played a role. Such cognitive bias within humans emphasizes certain types of bias when designing and developing a technology without being aware of it.

Additionally, human stereotypes can shape the design and objectives of AI systems, creating unintended consequences that limit the scope and fairness of technological solutions. Developers, influenced by their own cultural assumptions, may unconsciously encode these stereotypes into AI algorithms. For instance, models frequently display specific stereotypes associated with gender roles [11]. Some models may link cooking more strongly with women [12] or associate the term "CEO" with men [13]. Social networks were also under a heated debate as they are supposed to be a free platform for people to share their opinions respectfully. However, it is seen that some social networks are banning some opinions over the others and further promoting particular ones [14]. Such biased mechanisms of social networks are due to the biased owners or developers of social networks who usually force their opinions and views through the technology regulations.

Human tendencies to lie and distort reality further create challenges in achieving reliable human-AI collaboration. AI systems designed to process human-generated content, such as social media or survey responses, are often exposed to misinformation and deliberate manipulation [15]. This undermines the credibility of predictive models and decision-support systems, particularly in sensitive applications like healthcare and governance. Furthermore, some stakeholders with vested interests might deliberately falsify data or provide misleading inputs to influence the outcomes of AI systems. This can result in AI making decisions that align with the interests of a particular group, rather than the broader public good. O'Neil [16] mentioned that mathematical models and algorithms have attributes of opacity, scale, and destructive power. They work like a black box, with the process of generating results known to only a few people. However, these models are adapting from one domain to another and are being applied to the public. The poor and vulnerable groups become the victims.

The hallucination of AI and technology goes beyond the hallucination of human designers and developers to also cover human users of a technology. For instance, large language models (LLMs) are generally trained on extensive datasets collected from diverse online sources, tending to absorb toxic, offensive, misleading, stereotypical,

and other harmful or discriminatory content [17]. The Microsoft chatbot can exemplify how the bias of human users might be infused in AI and machines, leading to harmful impact. Specifically, Microsoft released in 2016 its Twitter (now called X) chatbot named Tay. The algorithm of Tay was developed to learn from other users' interactions on Twitter to get smarter and better answer users' queries. However, it is seen that in a short time, Tay started acting racist and making Nazi comments like "Hitler was right." The developers explained that during the algorithm learning phase from interactions, Tay inherited human biases and prejudices.

The human hallucination further goes beyond AI designers and developers to also cover experts. In a controversial incident at *NeurIPS*, one of the most popular conferences in the field of AI, a keynote presentation sparked significant backlash when it exemplified the misuse of AI with a particular nationality [18]. Responding to this incident, Jiao Sun, a Google DeepMind scientist, stated that "mitigating racial bias from LLMs is a lot easier than removing it from humans!" In research with collective voices discussing the opportunities and challenges of Generative AI (GenAI) in education, Bozkurt et al. [19] raised questions about whether researchers and developers are safeguarding equity and amplifying diverse voices—or reinforcing biases.

This raises the question of whether technology, originally designed to benefit humanity, may also exacerbate existing injustices. A prominent example of this is seen in facial recognition technology, where systems have been found to have higher error rates for people of color due to a lack of diverse representation in the training datasets [20]. AI systems are often viewed as objective or neutral tools, but in reality, they are not. Such inaccuracies are not only ethically problematic but also undermine public trust in AI systems.

Therefore, how do we expect to eliminate hallucination from a technology generally and AI particularly when the technology experts, designers, developers, and users (i.e., humans) are hallucinating? If we, researchers and practitioners, cannot maintain the highest standards of moral values, responsibility, and inclusion, how can we then develop ethical AI? Another key question is why to focus so much on AI and not that much on humans. AI is a support for reasoning, decision-making, processing, automation, and other functions. However, it is just that, a tool to support individuals, not to replace them. Thus, any AI hallucination is just an extension of human hallucination, the individual or group of individuals who created the database, algorithm, reasoning process, or collaborated in any other link in the chain of an AI-support tool, such as marketing, project design, or management. A failure to address these issues will cause technology to carry and amplify human biases, thereby reinforcing existing societal problems.

IV. Addressing Human Hallucination Is a Must To Mitigate AI Hallucination

On many occasions, the imperfection of human nature will cause or emphasize the hallucination of technology generally and AI particularly. To address this, it is important to first admit that we, humans and the technology developers and users, are not perfect. While several researchers highlighted, for instance, that eliminating bias from algorithms is easier than from humans [5], it is still crucial to put a lot of research efforts and investigations on humans to enhance ourselves (the technology creators and users) rather than on the technology. For instance, there should be more raising awareness about moral values, human responsibility, and accountability in technology (AI particularly) development, as well as the legal regulations and frameworks that developers need to respect in this context. So far, most of the debate is taking one strand, which is how to make ethical AI, while the question instead is how to make ethical

humans. If we simply spend time and efforts improving machines instead of ourselves, we might end up overpowered by them, and we become "slaves" of machines in the long run, just feeding them data and providing stronger computing powers.

Additionally, Carroll [21] stated that "a computer system does not itself elucidate the motivations that initiated its design, the user requirements it was intended to address, the discussions, debates and negotiations that determined its organization, the reasons for its particular features, the reasons against features it does not have, the weighing of tradeoffs, and so forth" (p.509). Therefore, it is crucial to rely on human-centered and human-in-the-loop approaches when designing a given software or hardware. This will allow capturing the real needs of users (not just the thoughts and visions of the designers and developers) who will be using the product and detecting any potential bias that might arise.

Based on Caroll's statement above, we ask ourselves, is it ethical to design a product to be used by everyone worldwide without having any or sufficient knowledge about each of the users? How can we expect that a product will be fair to millions of users, each of whom has a set of different interconnected variables (cultural, psychological, regional, religious, etc.) that makes them different from the others? Lewis and Rieman [22] stated that if you design something for everyone, it might well turn out to work for no one. This has been seen, for instance, in several GenAI tools that revealed discrimination and bias against several people. It is therefore important to ask if we want to design a product for a specific group of people that we really know about and make that product fair and effective for them or just design something for everyone, resulting in unfairness and maybe bias against some people. However, companies might not be in favor of the first as it will hinder their strategies of quick gains. Friedman and Nissenbaum [23] suggested that to minimize preexisting bias, "designers must not only scrutinize the design specifications, but must couple this scrutiny with a good understanding of relevant biases out in the world" (p.343). While admitting that identifying bias is very hard, they developed a framework to identify different types of biases that can be built into software and hardware, where bias is categorized into three main categories, namely pre-existing social bias, technical bias, and emergent social bias.

Moreover, following an inclusive thinking and design when developing AI-powered technologies is crucial. This implies that designers and developers must be open and inclusive in terms of considering various populaces in the moral creation and consumption of a technology. Such richness and diversity will allow mitigating any potential bias and discrimination.

Furthermore, it is important that more experimental testing with different people, contexts (economical, geographical, cultural, etc.), and needs is conducted before the final deployment of a technology. While most companies do not follow this as they think it is expensive and mainly rely on cost to make decisions related to a technology, cost-effectiveness only, unfortunately, does not predict the social and societal effects of that technology in the long run. In this context, Morningstar and Farmer [24] stated that "wherever possible, things that can be done within the framework of the experiential level should be. The result will be smoother operation and greater harmony among the user community. This admonition applies to both the technical and the sociological aspects of the system" (p.294).

V. Conclusions

Both humans and AI (technology generally) are hallucinating, and the first can cause or emphasize the latter. It is important therefore, when rethinking AI, to put more research, time, and effort on ourselves so that we can do better and improve as humans, especially morally. Particularly, it is much needed to conduct more research and investigation to understand the different types of human hallucination, how to detect them within a technology, and how they impact technology development and use. Identifying and understanding the link between human and AI hallucination will ultimately help to develop effective and safe AI-powered systems that could have some positive societal impact in the long run.

Considering open and inclusive approaches when designing and developing AI-powered systems is important. It is crucial to go beyond what designers, developers, and investors want to also consider the views and needs of users from different cultures, races, contexts, etc. This human-in-the-loop approach can help to mitigate the fundamental attribution error and create AI-powered systems that can potentially be used by everyone.

VI. Contents of This Monograph

This monograph is focused on the effects of culture on open science and artificial intelligence in education. The intersection of these key topics in the current panorama of higher education institutions and schools, along with any other educational level or setting, makes the monograph a milestone to understand better where we are and the next steps to take. Further, it sheds some light on a mid-term strategy so that the educational practitioners and facilitators go beyond the immediate response and focus on a n-step process into the future. With this vision in mind, the monograph collects a number of high-quality papers:

Pilicita-Garrido and Barra present how AI-supported sentiment analysis is vital to understand how cultural factors are instrumental for open science and artificial intelligence. They carry out a systematic review that shows pointers of benefits and challenges to boost an effective educational system.

Cotino-Arbelo et al. deal with youth expectations on working with Generative AI in higher education. They dig into the misconception and false expectations that popular views of artificial intelligence can project on youngsters. Through a thorough in-campus quantitative analysis, this research shows the level of misunderstanding in concepts and capabilities of AI.

Griffiths et al. present the European project GREAT, which focuses on citizen participation in climate change and environmental conflicts, through the use of an embedded survey in a mainstream game called SMITE. The results show that understanding and views on the core topic differ vastly amongst various age groups, genders, and education levels.

Denden and Abed introduce how Blockchain and AI facilitate the culture of sharing in an open science platform. The findings showed that the use of AI and Blockchain facilitates researchers and institutions working on open science environments to share more effectively and frequently.

Chen et al. show the practical use of ChatGPT as a tool to facilitate flipped classroom and how the students perceived that integration. More specifically, the class focused on enhancing students' understanding of traditional Chinese culture. This case study shows how to embed ChatGPT into daily classrooms as a tool for students and teachers.

Stracke et al. carry out an analysis of European policies on artificial intelligence in Europe. In a collective work with researchers from the European Union and the United Kingdom, the research analyses 15 policies on the topic, including comparisons amongst fundamental views and principles. Further, the study supports the combined use of AI in education with education about AI, which they call AI literacy.

References

[1] T. Hagendorff and S. Fabi, "Why we need biased AI: How including cognitive biases can enhance AI systems," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 36, no. 8, pp. 1885-1898, 2024.

[2] D. Dzhuhalyk, "Character.AI chatbot is accused of driving a teenager to suicide," Available: https://mezha.media/en/2024/10/24/character-ai-chatbot-is-accused-of-driving-a-teenager-to-suicide/

[3] A. Clark and M. Mahtani, "Google AI chatbot responds with a threatening message: "Human … Please die."," Available: https://www.cbsnews.com/news/google-ai-chatbot-threatening-message-human-please-die/

[4] N. Maleki, B. Padmanabhan, and K. Dutta, "AI hallucinations: a misnomer worth clarifying," *in 2024 IEEE Conference on Artificial Intelligence (CAI)*, 2024, pp. 133–138.

[5] E. Sengupta, D. Garg, T. Choudhury, and A. Aggarwal, "Techniques to eliminate human bias in machine learning," *in 2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*, November 2018, pp. 226–230.

[6] H. Ibrahim, F. Liu, Y. Zaki, and T. Rahwan, "Google Scholar is manipulatable," 2024, arXiv preprint arXiv:2402.04607.

[7] Y. Xu, M. Wang, K. Moty, and M. Rhodes, "How culture shapes the early development of essentialist beliefs," *Developmental Science*, vol. 28, no. 1, p. e13586, 2025.

[8] S. J. Watkins and C. Musselwhite, "Recognised cognitive biases: How far do they explain transport behaviour?," J*ournal of Transport & Health*, vol. 40, p. 101941, 2025.

[9] G. D. Baxter, E. F. Churchill, and F. E. Ritter, "Addressing the fundamental attribution error of design using the ABCS," *AIS SIGCHI Newsletter*, vol. 13, no. 1, pp. 76–77, 2014

[10] L. Ross, T. M. Amabile, and J. L. Steinmetz, "Social roles, social control, and biases in social-perception processes," *Journal of Personality and Social Psychology*, vol. 35, pp. 485–494, 1977.

[11] T. Bolukbasi, K. W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[12] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints," arXiv preprint arXiv:1707.09457, 2017.

[13] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach, "Women also snowboard: Overcoming bias in captioning models," *in Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 771–787.

[14] J, Guynn, "'You're the ultimate editor,' Twitter's Jack Dorsey and Facebook's Mark Zuckerberg accused of censoring conservatives." Available on: https://eu.usatoday.com/story/tech/2020/11/17/facebook-twitter-dorsey-zuckerberg-donald-trump-conservative-bias-antitrust/6317585002/

[15] G. Pennycook, A. Bear, E. T. Collins, and D. G. Rand, "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings," *Management Science*, vol. 66, no. 11, pp. 4944–4957, 2020.

[16] C. O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown, 2017.

[17] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of 'bias' in NLP," arXiv preprint arXiv:2005.14050, 2020.

[18] CTOL, "NeurIPS 2024 Sparks Controversy: MIT Professor's Remarks Ignite "Racism" Backlash Amid Chinese Researchers' Triumphs." Available on: https://www.ctol.digital/news/neurips-2024-controversy-mit-professor-remarks-chinese-researchers-triumphs/

[19] A. Bozkurt, J. Xiao, R. Farrow, J. Y. Bai, C. Nerantzi, S. Moore, and T. I. Asino (Eds.), "The manifesto for teaching and learning in a time of generative AI: A critical collective stance to better navigate the future," *Open Praxis*, vol. 16, no. 4, pp. 487–513, 2024.

[20] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," *in Conference on Fairness, Accountability and Transparency*, 2018, pp. 77–91.

[21] J. M. Carroll, "Human-computer interaction: psychology as a science of design," *International Journal of Human-Computer Studies*, vol. 46, pp. 501–522, 1997.

[22] C. Lewis and J. Rieman, Task-Centered User Interface Design: A Practical Introduction, 1993. Published as shareware. Available from: https://hcibib.org/tcuid/tcuid.pdf.

[23] B. Friedman and H. Nissenbaum, "Bias in computer systems," *ACM Transactions on Information Systems*, vol. 14, no. 3, pp. 330-347, 1996.

[24] C. O. Morningstar and F. R. Farmer, "The lessons of Lucasfilm's Habitat," in B. Michael, Ed., Cyberspace: The First Steps. Cambridge, MA: MIT Press, 1991.

Ahmed Tlili

Ahmed Tlili is an Associate Professor at Beijing Normal University, China, Adjunct Associate Professor at An-Najah National University, Palestine, and a Visiting Professor at Universidad Internacional de La Rioja (UNIR), Spain. He is the Co-Director of the OER Lab at the Smart Learning Institute of Beijing Normal University (SLIBNU), China. He serves as the Editor of Springer Series *Future Education and Learning Spaces*, and the Deputy-Editor-in-Chief of *Smart Learning Environments*. Prof. Tlili is also an expert at the Arab League Educational, Cultural and Scientific Organization (ALECSO). He has edited several special issues in several journals. He has also published several books, as well as academic papers in international referred journals and conferences. He has been awarded the Martin Wolpers 2021 Prize by the Research Institute for Innovation and Technology in Education (UNIR iTED) in recognition of excellence in research, education and significant impact on society. He also has been awarded the IEEE TCLT Early Career Researcher Award in Learning Technologies for 2020. He has been listed in the Stanford/Elsevier top 2% influential scientists worldwide for 2024.

Daniel Burgos

Daniel Burgos is a full professor of Technology for education and communication and vice-rector for international research at the Universidad Internacional de La Rioja (UNIR). He holds a UNESCO Chair on eLearning. He is the Director of the Research Institute for Innovation and Technology in Education (UNIR iTED, http://ited.unir.net). Also, he is the President of MIU City University Miami, USA. He has implemented more than 80 European and worldwide R&D projects and published more than 270 scientific papers, 60 books and special issues, and 12 patents. He is a full professor at An-Najah National University (Palestine), an adjunct professor at Universidad Nacional de Colombia (UNAL, Colombia), an extraordinary professor at North-West University (South Africa), a visiting professor at the China National Engineering Research Center for Cyberlearning Intelligent Technology (CIT Research Center, China); and a research fellow at INTI International University (Malaysia). He works as a consultant for the United Nations (UNECE), the United Nations University (UNU-FLORES), ICESCO, the European Commission and Parliament, and the Russian Academy of Science. He holds 13 PhD degrees and doctorates, including Computer Science and Education. ORCID: 0000-0003-0498-1101.

Sentiment Analysis With Transformers Applied to Education: Systematic Review

Anabel Pilicita-Garrido, Enrique Barra *

Departamento de Ingeniería de Sistemas Telemáticos, Universidad Politécnica de Madrid, Madrid (Spain)

* Corresponding author: a.pilicita@alumnos.upm.es (A. Pilicita-Garrido), enrique.barra@upm.es (E. Barra Arias)

Abstract

Sentiment analysis, empowered by artificial intelligence, can play a critical role in assessing the impact of cultural factors on the advancement of Open Science and artificial intelligence. Additionally, it can offer valuable insights into the open data gathered within educational contexts. This article presents a systematic review of the use of Transformers models in sentiment analysis in education. A systematic review approach was used to analyze 41 articles from recognized digital databases. The results of the review provide a comprehensive understanding of previous research related to the use of Transformers models in education for the task of sentiment analysis, their benefits, challenges, as well as future areas of research that can lay the foundation for a more sustainable and effective education system.

I. Introduction

SENTIMENT analysis, also known as opinion mining, is one of the most well-known tasks of Natural Language Processing (hereinafter, NLP). This task identifies how sentiments are expressed in words, sentences, or writings toward a particular topic [1]. In essence, it involves finding out the attitudes, opinions, preferences, and sentiments of users by researching, analyzing, and mining subjective texts [2]. In educational environments, students' emotions play a crucial role in the learning process because they can enhance or undermine their ability to learn or remember what they have learned [3]. Opinion mining is a multidisciplinary field that can be applied to different educational domain challenges, such as course evaluation, understanding student participation, educational infrastructure constraints, and educational policy decision-making [4]. In this sense, when information is extracted from reviews left by students, it can improve teaching and learning practices [5]. Yan et al. [6] and Du [7] conducted studies to analyze student feedback, revealing several critical factors that influence student satisfaction in virtual learning environments. These factors include course content, technical elements, difficulty level, instructor proficiency, video resources, course organization, and workload.

Culture is transmitted between generations through text, images, audio, video, or traditions that generate collective memory [8], [9]. Sentiment analysis is an emerging tool for analyzing cultural phenomena. Culturally, language choice affects moral decisions, suggesting the importance of language in message transmission [10]. Similarly, Lennox et al. [11] consider that sentiment analyses culturally provide better data to address the human side of conservation. In this sense, the opinion or behavior of humans in relation to certain topics of interest varies according to how feelings are expressed, perceived, and interpreted in different cultures. In fact, culture greatly influences social behavior, communication, cognitive processes, and pedagogical technology [12]. Therefore, sentiment analysis can be crucial in determining the influence of culture on Open Science and artificial intelligence, providing a solid understanding of the open data collected in educational environments and contexts.

In 2017, a new Transformers architecture was proposed, applying parallel computing and transfer learning with a self-attention mechanism [13]. This architecture is simple and has shown that it was possible to design this type of network with good results in NLP tasks such as sentiment analysis with a set of multiple sequential attention layers [14]. Nowadays, there is little research that mentions Transformers architecture in a systematic literature review (SLR) on sentiment analysis applied to education. Previous reviews of the literature identified machine learning techniques and algorithms that are prevalent in sentiment analysis in education. A study by Oghu et al. [15] examined 59 relevant papers and the authors identified five common techniques that are mainly investigated for sentiment analysis in education and the prevailing supervised machine learning algorithms. Shaik et al. [16] conducted research on sentiment analysis using educational data and found that educational institutions have invested heavily in creating sentiment analysis tools and applications based on student opinion analysis. In the paper, the authors explored the challenges of sentiment analysis, such as multipolarity, polysemy, negation words, and opinion spam detection. These two studies only

mention Transformers models without expanding their scope of application.

Given the lack of a literature review on the current state of the art of sentiment analysis with Transformers architecture in education, it is important to establish the current state of the art of this topic and determine how it can influence culture and Open Science. Additionally, it is also important to determine challenges and future applications that can improve the educational field. This research seeks to address the following research questions (RQ).

- RQ1: What is the current state of the art in sentiment analysis research with Transformers applied to education?
- RQ2: What are the main benefits of applying sentiment analysis with Transformers in education?
- RQ3: What are the main challenges in applying sentiment analysis with Transformers in education?
- RQ4: What are the possible future areas of research around sentiment analysis with Transformers in education?

This paper presents a systematic review of sentiment analysis with the Transformers architecture in the educational setting. The main aim of this investigation is to analyze and present a general summary of research related to this topic. This is necessary to provide updates on the state of the art, identify well-researched areas, reveal lagging areas that need further research, and understand the main trends in this field. Finally, to provide a useful summary of current knowledge in a field of study, and, consequently, possible research directions.

II. Overview of Transformers Technology

Although the human-machine interaction may seem simple in theory, it is complex and difficult in practice. In fact, deep learning models have addressed challenges in NLP tasks that present significant results [17]. However, in 2017 the NLP field was revolutionized with the origin of the Transformers models. Transformers have a deep learning architecture that differentiates the importance of each input sequence based on an attention mechanism. Therefore, words are detected by this mechanism, which means that the data are not necessarily processed in an ordered way. Furthermore, the architecture is based on an encoder-decoder structure. The encoder layer processes the input layer by layer, while the decoder layer does the same in the opposite direction [13]. Thus, the encoder layer encodes the input, while the self-attention mechanism assigns weights to each input according to its relevance [13], [18]. Due to the attentional mechanisms, Transformers allow for greater parallelization during training and can be trained faster, improving their performance on NLP tasks. However, the main disadvantage of the Transformers architecture is the high computational cost when training and testing models.

In the educational field, Transformers have been applied in some NLP tasks to contribute to teaching and learning processes by automating repetitive tasks. For example, different models have been applied to the automatic grading of essays [19], the automatic grading of multiple-choice tests [20], and even the automatic grading of short answers [21]. Text generation is another task that has been applied in the educational domain for the automatic summarization of high-quality texts [22] or for creating story endings [23]. This task has also been related to the paraphrasing of original texts with different expressions [24]. Finally, models have also been used to create multiple-choice questions for assessments [25] or to automatically classify the students' responses for further evaluation [26], [27]. These studies suggest that the application of natural language processing can contribute to Open Science by ensuring that scientific knowledge is accessible and that the production of that knowledge itself is inclusive, efficient, equitable, and sustainable.

Sentiment analysis, often referred to as opinion mining, is one of the most widely used NLP applications for identifying human intentions based on opinions [16]. While these terms are commonly used interchangeably in the academic literature, they are not synonymous. However, for the purposes of this article, they will be treated as such. Chats, remarks, opinions, or comments with emotional valuations are part of public opinion on the Internet. In general, any kind of topic is discussed on social networks, forums, news, or other types of digital spaces. In recent years, Transformers models have been applied for sentiment classification [28]. Additionally, an enormous amount of text data often needs to be automatically reviewed, classified, and filtered malicious categories such as hate speech, fake news, or spam [16]. In this case, efficient emotion recognition has aroused great research interest in presenting methodological proposals that focus on stable and accurate results [29], or frameworks specialized in optimizing tasks [30]. Similarly, other research focused on the quality of the explanatory text to improve user confidence and satisfaction by generating recommendations [31] and a public opinion sentiment analysis method to improve the efficiency of sentiment trend analysis [32].

It is important to consider in sentiment analysis the recognition of textual emotion because information may be limited or ambiguous. Different approaches have been proposed for identifying emotions. Ekman et al. [33] indicated the existence of six emotions: happiness, sadness, anger, fear, disgust, and surprise. Izard et al. [34] considered twelve emotions: interest, joy, surprise, sadness, anger, disgust, contempt, self-hostility, fear, shame, shyness, and guilt. However, Ekman's approach is one of the most widely applied approaches in natural language processing. Additionally, Bruna et al. [35] established categorical emotion models in which the textual recognition of emotions is based on the idea of discrete emotion theory, in which the categorical classification is simpler and consists of deciding whether the emotion is positive or negative. In general, to evaluate emotion, lexicons with appropriate emotional content are used, or a classifier is trained with a well-annotated database indicating the nature of the emotions.

Currently, different models based on the Transformers architecture can be found such as BERT (Bidirectional Encoder Representations from Transformers) [36], RoBERTa [37], AlBERT [38], XLNet [39], DistilBERT [40], Reformer [41], GPT-2 [42] and GPT-3 [43]. One of the models that has achieved excellent results in the NLP field is BERT with its respective variants [44], [45]. This model is present in most research aimed at classifying opinion aspect sequences [46], [47], high-quality word detection [48], and significant feature detection for classifying fake news and real news [49], [50]. Therefore, studies have focused on improving the performance [51], [52], stability [53], [54], efficiency [18], [55] and robustness against missing data [56] - [58] of Transformers models intended for sentiment classification. Finally, automatic term extraction is an important task in sentiment analysis, this is the case for a text in which keywords are identified to improve sentiment prediction [56].

III. Methodology

Systematic reviews of the literature are classified into domain-based reviews, theory-based reviews, and method-based reviews [59]. The methodological approach taken in this study is Callahan´s method [60] and this method belongs to the category of domain-based reviews. Callahan´s guidelines consist of a systematic literature review with the 6W framework (Who, When, Where, hoW, What, and Why). Therefore, this paper applied Callahan´s method to present a review of existing literature on the use of sentiment analysis in education with Transformers and the key information on the method is summarized in Table I.

TABLE I. Key Information for the Systematic Literature Review Adapted From Callahan's 6W Framework [60]

Who conducted the review?	The authors of this paper
When were the data collected?	From October 2023 to December 2023
Where were the data collected?	Six electronic databases (Scopus, ScienceDirect, IEEE Xplore, SpringerLink, Taylor & Francis, and MDPI) were searched for articles in peer-reviewed, scholarly journals and conferences
How were the data found?	The identification of literature relevant to the topic involves considering research questions that establish a search for articles on uses, benefits, challenges, or future applications of the Transformers architecture in sentiment analysis in the field of education. Subsequently, Keywords extracted from research questions are given below: "education", "student", "MOOC", "learning", "teaching", "sentiment analysis", "opinion mining", "Transformer" and "pretrain model". The search strings were then defined to be used in the bibliographic databases of Scopus, ScienceDirect, IEEE Xplore, SpringerLink, Taylor & Francis, and MDPI
What was found?	A final data set of 41 articles
Why were certain works included?	Search words found in title, abstract, or keywords; English; explicitly on sentiment analysis applied in education with Transformers models.

A. Data Collection

Data for this study were collected between October and December 2023. To establish the search interval, the emergence of Transformers in 2017 was taken as a reference [13]. Consequently, the research was restricted to literature published from 2017 to December 2023. Data included literature that met the following selection criteria: published journal or conference written in English, mention of sentiment analysis with Transformers applied to education in title, abstract, or keywords.

A pilot search in Scopus showed that a search using the term "sentiment analysis" would yield more than 42.954 articles. In contrast, a more specific search with Boolean operators was performed including title, abstract, and keywords. Therefore, the search keywords were "sentiment" AND "analysis", AND "transformers", AND "education". In this case, there were six publications returned. It was therefore determined that the first search led to many publications that referred to an unmanageable number of articles, and the second search was evidently too restrictive. As a result, the data were collected in two phases, a process illustrated in Fig. 1.
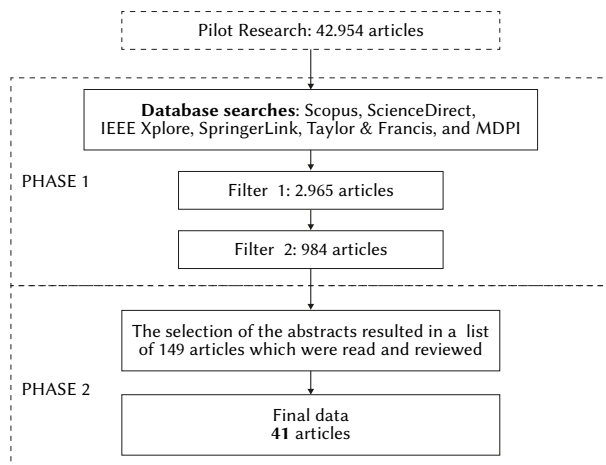
Fig. 1. Overview of the search and review process.

B. The First Search Phase

Education is a vast field and covers different topics. Sentiment analysis can also be referred to as opinion mining. Due to this situation, synonyms and other useful keywords were used for advanced search. This is presented in Table II.

TABLE II. Keywords With Their Respective Synonyms

Keyword	Synonym
Sentiment analysis	opinion mining
Transformers	pretrain model
Education	training instruction teaching learning academic learning study tutoring e-learning

The next step was to apply some of the synonyms in searches with Boolean operators OR and AND. However, the first results were associated with other types of research. For example, the term "learning" was associated with the term "machine learning", which refers to models other than Transformers, but is equally related to the field of artificial intelligence. In this case, it was decided to keep the term "Transformer" or "pretrain model" and thus limit somewhat the inclusion of other types of research. One situation that could be observed in the development of the advanced search was that some studies were left out by not considering the words associated with the analysis of student sentiment or the analysis of online course comments. Therefore, to minimize the risk of omitting relevant publications, the searches were customized to include useful words such as "student" and "MOOC". Next, Table III shows the keywords and Boolean operators applied in the advanced search of this first phase. The database search engines used were Scopus, ScienceDirect, IEEE Xplore, SpringerLink, Taylor & Francis, and MDPI. In this case, publications were extracted in the form of research articles, and the search was performed on titles, abstracts, and keywords.

In Table III, a "filter 1" column is displayed, which corresponds to the application of the advanced search with 2.965 results. In this first filter, some considerations were taken into account. For example, the search words changed in two cases in the ScienceDirect database where the word "academic" was not included because the database limited the number of words. Therefore, tests were carried out to verify that removing a word did not alter the search result set. The second case was in the MDPI database where the search was changed because the results became restrictive as more words were added. Therefore, the search was continued with "sentiment analysis" OR "opinion mining". Finally, filter 2 considered only English-language open-access articles published between 2017 and 2023, obtaining 984 results in this first phase.

C. The Second Search Phase

In the second phase of the process, the 984 articles derived from the initial phase were reviewed. The review of the abstracts was based on the following inclusion and exclusion criteria.

1. Inclusion Criteria

Articles that (a) focused exclusively on sentiment analysis applied to education using the Transformers architecture and (b) were original, peer-reviewed publications in open-access journals or conferences, were included.

TABLE III. Advanced Database Search

Database	Advanced Search	Returned Articles	
		Filter 1	Filter 2
IEEE Xplore	"All Metadata": education OR "All Metadata": student OR "All Metadata": MOOC OR "All Metadata": learning OR "All Metadata": academic OR "All "Metadata": teaching) AND ("All Metadata": "sentiment analysis" OR "All Metadata": "opinion mining") AND ("All Metadata": "Transformer" OR "All Metadata": "pretrain model"	603	80
ScienceDirect	(education OR student OR MOOC OR learning OR teaching) AND ("sentiment analysis" OR "opinion mining") AND ("Transformer" OR "pretrain model")	63	25
Scopus	(education OR student OR "MOOC" OR learning OR academic OR teaching) AND ("sentiment analysis" OR "opinion mining") AND ("Transformer" OR "pretrain model")	838	255
SpringerLink	("education" OR "student" OR MOOC OR "learning" OR "academic" OR "teaching") AND ("sentiment analysis" OR "opinion mining") AND ("Transformer" OR "pretrain model")	970	217
Taylor and Francis Online	("education" OR "student" OR MOOC OR "learning" OR "academic" OR "teaching") AND ("sentiment analysis" OR "opinion mining") AND ("Transformer" OR "pretrain model")	100	43
MDPI	("sentiment analysis" OR "opinion mining")	391	364
TOTAL		**2.965**	**984**

2. Exclusion Criteria

Articles that (a) were not related to the Transformers architecture, (b) did not apply the Transformers architecture in the educational setting, and (c) only showed the performance results of Transformers models, were excluded. In addition, non-journal or non-conference peer-reviewed articles were also excluded.

The evaluation of the abstracts led to the identification of a list of 149 articles which were subjected to a thorough reading and detailed review.

As a result, articles were selected that demonstrated the practical value of analyzing feelings using Transformer models in education, with the aim of understanding how this technology is being implemented in the field and what it contributes. Subsequently, the articles that deviated from the scope of the research were excluded, resulting in a final list of 41 selected articles.

IV. Results

This section presents the results obtained in the present work. The 41 articles were classified into three categories. Category A corresponds to the research that collected the analysis data in online educational environments. Category B has the works that collected the analysis data from social networks or platforms such as Google Play, and category C finds the articles that obtained the analysis data from different media such as audio or surveys. After ranking the 41 items, the percentages occupied by each of the labeled categories were: category A (46%), category B (32%), and category C (22%). The coding scheme was employed to obtain information from the studies, based on Table IV.

TABLE IV. Categories and Items Included in the Coding Scheme of This Systematic Review

Coding Domine	Items
Source information	Year of publication Author
Study Quality	Source of data
Setting characteristics	Research objective Data collection spaces

The results obtained provide a comprehensive overview of the evolution and trends in sentiment analysis research with Transformers in education. It is important to emphasize that the presented works start from the year 2019. These findings may be associated with the scientific publication cycle, wherein the research process, from conceptualization to publication, takes time. Therefore, many projects could have begun in the years preceding 2019, but their results were disclosed in conferences or scientific journals in that specific year. In the following sections, the papers will be presented according to the established categories, and different research based on Transformers models will be exposed considering that each of them has its respective scope according to the objective set. The compilation of all studies can be visualized in Table V.

A. Data Extracted From Online Educational Environments

The first section will examine research works that have collected opinions, comments, chats, or reviews left by students in online educational environments, in such a way that they go through a sentiment analysis process that contributes to improving the educational processes of teaching and learning. A growing body of literature has investigated e-learning, a teaching and learning system based on the use of Information and Communication Technologies (ICT) [61]. More recent attention has focused on applying a Transformer-based model that performs automatic sentiment prediction. Therefore, it is important to mention research that as a result evaluates the accuracy of a model, other research that develops a deep analysis of emotions, and more findings that are related to educational improvement.

In recent years, there has been an increasing amount of literature on Massive Open Online Courses (MOOCs). One of the most well-known teaching systems is a type of online course designed to be accessible to many participants around the world. According to Zhou et al. [62] reviews left by students in these courses are taken to accurately classify each of them or to evaluate the quality of the courses according to the sentiments detected. This view is supported by Yan et al. [63] and Du [7], who in their research extracted reviews posted by students on MOOC courses and determined that course content, technical content, degree of difficulty, teaching staff level, video provided, course schedule, and homework load are the main factors affecting student satisfaction in these virtual learning environments. Together, these studies show that these factors help to identify deficiencies in the course to act and improve the quality of the course. Moreover, Jatain et al. [64] analyzed student feedback from Coursera to capture aspects that influence the popularity of the course, although the study focused on assessing the pressure on the BERT model´s prediction of sentiment.

The authors of two studies [65], [66] conducted a sentiment analysis of MOOC course reviews to compare Transformers models with other deep learning methods emphasizing the superior performance of models such as BERT or RoBERTa. Similarly,

Marfani et al. [67] analyzed four Coursera courses and extracted opinions and comments from students for sentiment analysis to help identify their shortcomings and improve their courses. In the case of this research, the aim was to find the model that worked best to identify the key aspects of sentiment classification by basing the experiments on real student data. The accuracy rate obtained in this research with BERT was outstanding. A broader perspective was adopted by Pan et al. [68] who applied the BERT model. In this case, after the classification of feelings from the comments left by the students, the results obtained were analyzed in depth. Therefore, it was determined that academic emotions improved significantly in the first and second periods of the course and tended to be stable in the second and third periods of the course. In the same vein, a pre-trained ALBERT model was also applied to extract key information from comments left in MOOCs courses and it was determined that the model overcame the problem of the traditional sentiment analysis method that cannot distinguish the different meanings of the same word in different contexts [69]. Research has also proposed new models according to language needs. Conversely, Min et al. [70] proposed in their work to extract reviews of Chinese online courses using an ALBERT model by analyzing a small amount of labeled data to choose courses with better reviews.

In addition to reviews posted on MOOCs, forums are one of the means of communication between teachers and students, although it is difficult to distinguish messages that require prompt intervention considering a priority level. In this sense, sentiment analysis helps teachers prioritize responses to questions left on forums promptly [71]. Liu et al. [72] collected 8.867 student posts in a forum and identified the interactive relationship between emotional and cognitive engagement in students' learning process. In other words, positive or confusing emotions have been determined to contribute more to high-level cognition than negative emotions. Cognitive screening was also conducted in MOOCs to identify unlabeled messages from two MOOC courses and determine the cognitive presence of learners. These results provide valuable information on the effectiveness of pre-training in large-scale multidisciplinary discussion data [73]. These articles reveal the relationship between cognitive aspects and the opinions left in the forums.

Recently, researchers have paid attention to more sentiments. As is the case of Alkaabi et al. [75] who used messages left on online learning platforms to classify students' emotions as positive, negative, or neutral. They then managed to extract dominant negative sentiments such as anger, disgust, fear, and sadness from students. In this sense, chats between teachers and students have also made it possible to analyze feelings on these platforms to help teachers improve their teaching methods [79]. In these works, the pre-trained BERT model was applied to perform sentiment classification, achieving high accuracy in emotion prediction.

So far, all research has presented data collected exclusively from MOOCs. However, studies have been written that perform sentiment analysis with multimodal data that are taken from MOOCs and other different educational spaces. In a sentiment analysis, Qu et al. [74] classified student behavioral data collected from the course evaluation system and the academic management system, including textual information from students' comments on the course. Another author, Dyulicheva [77], investigated mathematics MOOCs and student reviews. In sentiment analysis, deep learning was applied to identify some clusters of various negative emotions related to students past bad mathematics experiences. As a result, students' emotional states associated with math phobia represent substantial barriers to learning mathematics and acquiring basic mathematical skills. Together, these studies outline the high performance of the Transformer BERT model with multimodal data.

An interesting aspect of the research focused on sentiment analysis in this category is proposals with new models based on Transformers. For example, BERT model can be modified or fine-tuned to train it with different data. More recent attention has focused on comparing Transformer models and determining their performance. In this sense, among studies that compared Transformers models with other machine learning models, the superiority of Transformers models in sentiment analysis prediction was highlighted due to the inherent ability to capture complex patterns and long-range relationships in text sequences.

Overall, these studies indicate that sentiment analysis focused on the educational field can improve the teaching and learning process, especially on online learning platforms. One possible implication of this is that there is great value in the information found in learning systems because it is objective and didactic data that are useful for school management and can improve course development. Students' academic emotions are important for academic performance and contribute greatly to educational success because they generate new problem-solving proposals that manifest themselves in online educational environments.

B. Data Extracted From Social Media

The second section will examine research that has collected feedback left by students outside of online educational environments, such as social networks or other platforms. More recent attention has focused on social networks because they have a large and diverse population, so people can express their opinions on any topic daily. Collectively, comments are open to all audiences, and students use this medium to express feelings or opinions also in an educational context [88]. Some studies analyze comments on topics that can influence students' decision-making for their vocational training. For example, Fouad et al. [91] used more than 250.000 tweets in sentiment analysis with the BERT model to analyze perceptions of women in STEM (Science, Technology, Engineering, and Mathematics) fields. Many of the opinions were positive about women's entry into these fields, and consequently, in this sentiment analysis, it was determined that the positive aspects may encourage women to enroll in higher education careers related to them. Another study also analyzed around 3.542 Reddit forum posts about the Radiology career with the RoBERTa Transformer and the perception of the Radiology career was positive for the students [82]. Considering both results, the mostly positive comments may encourage students to make enrollment decisions.

However, in contrast to earlier findings, there have been several studies that reveal critical opinions about educational topics. The research by Zhou and Mou [86] found that the expectations associated with online learners are difficult to meet, as they expect students to be self-disciplined and self-regulated. It is expressed that students must stare at the screen for a long time due to long online sessions; this situation was associated with keywords such as "drowsy" and "anxious" in the sentiment classification. In the same vein, another study analyzed student tweets, in which it was found that students may suffer from stress in this modality [83]. There is no doubt that online education is an omnipresent force that can meet emergent needs during unexpected events to provide continuity in education, but studies have also exposed critical views. Therefore, the authors mention the need to diversify online teaching and learning activities to maximize student attention. Prasad et al. [80] address the challenge of manually annotating large volumes of data, they proposed a machine learning method that uses the sentiment 140 dataset as a training set to automate the process of tagging student tweets on social networks. They conclude that this method can be effectively applied to label any qualitative data.

TABLE V. THE USE OF TRANSFORMERS IN SENTIMENT ANALYSIS IN EDUCATION

S/N	Article Title	Authors	Publication Year	Category
1	Sentiment Analysis of MOOC Reviews Based On Capsule Network [69]	Liu et al.	2021	A
2	A Shallow BERT-CNN Model for Sentiment Analysis on MOOCs Comments [65]	Li et al.	2019	A
3	Analysis of Learners' Sentiments on MOOC Forums using Natural Language Processing Techniques [67]	Marfani et al.	2022	A
4	Can We Predict Student Performance Based on Tabular and Textual Data? [74]	Qu et al.	2022	A
5	Online Course Quality Evaluation Based on BERT [62]	Zhou et al.	2020	A
6	Bi-GRU Urgent Classification for MOOC Discussion Forums Based on BERT [71]	Khodeir	2021	A
7	MOOC-BERT: Automatically Identifying Learner Cognitive Presence From MOOC Discussion Data [73]	Liu et al.	2023	A
8	An exploration of the causal factors making an online course content popular & engaging [64]	Jatain et al.	2023	A
9	Automated detection of emotional and cognitive engagement in MOOC discussions to predict learning achievement [72]	Liu et al.	2022	A
10	Deep learning for opinion mining and topic classification of course reviews [66]	Koufakou	2023	A
11	Sentiment Analysis and Topic Mining Using a Novel Deep Attention-Based Parallel Dual-Channel Model for Online Course Reviews [63]	Yan et al.	2023	A
12	Detecting Emotions behind the Screen [75]	Alkaabi et al.	2022	A
13	Cross-Domain Polarity Models to Evaluate User eXperience in E-learning [76]	Sanchis-Font et al.	2021	A
14	Learning Analytics in MOOCs as an Instrument for Measuring Math Anxiety [77]	Dyulicheva	2021	A
15	A Small amount of Labeled Data Chinese Online Course Review Target Extraction via ALBERT-IDCNN-CRF Model [70]	Min et al.	2020	A
16	Case Study: Predicting Students Objectivity in Self-evaluation Responses Using Bert Single-Label and Multi-Label Fine-Tuned Deep-Learning Models [78]	Nikolovski Vlatko and Kitanovs	2020	A
17	Sentiment Analysis of Comment Texts on Online Courses Based on Hierarchical Attention Mechanism [79]	Su et al.	2023	A
18	Research on the factors influencing the learner satisfaction of MOOCs [7]	Du	2023	A
19	Are students happier the more they learn? – Research on the influence of course progress on academic emotion in online learning [68]	Pan et al.	2022	A
20	Supervised Sentiment Analysis of Indirect Qualitative Student Feedback for Unbiased Opinion Mining [80]	Prasad et al.	2023	B
21	Sentiment Analysis of Students' Feedback on E-Learning Using a Hybrid Fuzzy Model [81]	Alzaid et al.	2023	B
22	Broadening the Understanding of Medical Students' Discussion of Radiology Online: A Social Listening Study of Reddit [82]	Hameed et al.	2023	B
23	Sentiment Analysis of Stress Among the Students Amidst the Covid Pandemic Using Global Tweets [83]	Jyothsna et al.	2023	B
24	Sentiment Analysis of Tweets on Online Education during COVID-19 [84]	Yldrm Elif and Yazgan	2023	B
25	Sentiment Analysis and Topic Modeling on Tweets about Online Education during COVID-19 [85]	Mujahid et al.	2021	B
26	Tracking public opinion about online education over COVID-19 in China [86]	Zhou et al.	2022	B
27	The Ivory Tower Lost: How College Students Respond Differently than the General Public to the COVID-19 Pandemic [87]	Duong et al.	2020	B
28	Research on the Method of Identifying Students' Online Emotion Based on ALBERT [88]	Ren et al.	2021	B
29	Sentiment Analysis of Code-mixed Social Media Data on Philippine UAQTE using Fine-tuned mBERT Model [89]	Maceda et al.	2023	B
30	Analysing user reviews of interactive educational apps: a sentiment analysis approach [90]	Mondal et al.	2022	B
31	Sentiment Analysis for Women in STEM using Twitter and Transfer Learning Models [91]	Fouad et al.	2023	B
32	Students' preferences with university teaching practices: analysis of testimonials with artificial intelligence [92]	Álvarez-Álvarez et al.	2023	B
33	A Multi-Modal Convolutional Neural Network Model for Intelligent Analysis of the Influence of Music Genres on Children's Emotions [93]	Qian and Chen	2022	C
34	Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system [94]	Javaid et al.	2023	C
35	Sentiment Analysis of Using ChatGPT in Education [95]	Tubishat et al.	2023	C
36	Online teaching emotion analysis based on GRU and nonlinear transformer algorithm [96]	Ding	2023	C
37	Arabic Sentiment Analysis for Student Evaluation using Machine Learning and the AraBERT Transformer [97]	Munshi et al.	2023	C
38	Analysis of the Sentiment in the Evaluation Texts of University Students by Means of the Concept of Flexible Management [98]	Zhu	2023	C
39	Aspect and Sentiment Classification Mechanisms of Student After-Class Self-Evaluated Comments: Investigation on Nonsense Data, Feature Extraction, and Classification Models [99]	Chou et al.	2023	C
40	Multilabel Classification of Student Feedback Data Using BERT and Machine Learning Methods [100]	Setiawan et al.	2023	C
41	Towards Application of Speech Analysis in Predicting Learners' Performance [101]	Chowdary Attota et al.	2022	C

Turning to the opinions of students on the topic of the pandemic on social media, Duong et al. [87] in their research analyzed 73.787 tweets from 12.776 university Twitter followers, and only tweets related to the pandemic were considered. They found that university students were significantly more negative due to the spread of COVID-19 and were observed with racist comments toward the Asian community. It is important to mention that in this study the data were collected in the year 2020 during the pandemic, which is why these results were observed. In the same vein, Mujahid et al. developed two studies focusing on sentiment analysis of students' emotions about the pandemic using e-learning Twitter data [85] and distance education data to classify sentiments into positive, negative, and neutral [84].

Studies on sentiment analysis have also focused on languages other than English. For example, a topic of interest to students is access to universities in the Philippines. In this case, around 13.332 student comments were collected on Twitter, Facebook, and YouTube. Then the multilingual Bidirectional Encoder Representations from Transformers (mBERT) model was applied, achieving 80% accuracy, and determining that students view third-level education positively but also revealed concerns about delays in grants or alleged misuse of funds [89]. Additionally, a study analyzing the sentiment of student comments on e-learning in Saudi Arabia found that comments were ambiguous, and opinions were unclear [81]. This study suggests that the opinion of a personal text differs depending on the context or setting in which it is expressed, and in social networks, informal language can also be a drawback when analyzing sentiment. These findings contribute to the understanding that the application of Transformers models extends to other languages while maintaining strong performance in sentiment prediction.

On the other hand, it is feasible to analyze the public opinion of students themselves on platforms such as the Google Play Store. Mondal et al. [90] analyzed over one million reviews of 800 Augmented Reality (AR) and Virtual Reality (VR) apps with an educational focus. The results suggested that educational applications that did not incorporate AR or VR received higher user satisfaction than applications that incorporated these technologies. In the same way, the study by Sanchis et al. [76] analyzed users' feelings based on preferences, behaviors, and achievements before, during, and after interaction with virtual learning environments. Likewise, Nikolovski et al. dealt with student opinions about teachers obtained objectively for further analysis using these educational software [78]. These results could emphasize categorizing user opinions within a technological environment and hold great potential for guiding educational software developers in refining the design of programs, applications, or virtual learning platforms. By gaining insights into user sentiments, developers can better align their innovations with the specific functionalities desired by users.

These investigations agree with the findings of other studies located in the first section, in which opinions or comments allow valuable information to be extracted and thus take action. Likewise, some studies [80], [81] expressed the belief that one of the drawbacks detected in this research is the lack of context in opinion detection, as it decreases the likelihood of correct sentiment detection, due to the use of jargon, youth code, or various forms of expression used by students. Therefore, pre-processing and cleaning the data contributes to a better sentiment classification process with Transformers models. Sentiment analysis with Transformers in different languages is still scarce, and the application of Transformers in research focuses on fitting pre-trained models and evaluating their behavior in terms of performance metrics.

C. Data Extracted From Different Educational Spaces.

This section presents different scenarios that have been collected for sentiment analysis. For example, Setiawan et al. [100] collected student feedback from the comments left by students in mentoring sessions with their tutors. This study aimed to find out the students' queries for the university departments to provide answers to ensure satisfactory delivery of services to the students. Therefore, the BERT model was applied to classify students' comments in mentoring sessions as positive or negative, achieving an accuracy of 82% [100]. A broader perspective has been adopted by Zhu [98] who used university students' teaching evaluation texts for a sentiment analysis based on BERT. The objective of this study was to understand and more accurately analyze the teaching evaluation texts of university students by applying sentiment analysis to explore the deep semantic information of the texts.

On the other hand, in the study by Munshi et al. [97] a dataset was created from student surveys. At the end of the manual collection, 1.044 student responses were obtained, reaching 3.472 comments related to preferences about a subject. In this data analysis, a BERT-based model called AraBERT was applied to classify feelings into positive, negative, and neutral, achieving an accuracy of 82%. Similarly, in another investigation [99], students were asked to write their self-assessed comments in a system after each class. In total, 1.640 anonymous comments were collected. The applied model was BERT, with an accuracy of 93%, and the classification was made into three categories: positive, negative, and neutral. An interesting aspect of this study is that for the classification of the three categories mentioned above, seven aspects: interest, gain, positivity, speed, difficulty, teacher in general, and others were applied. Chowdary Attota et al. [101] collected students' moods through audio recordings in class while discussing the course topic in teams. Pre-trained models with transformational linguistic models were applied for automated audio data transcription and conversion achieving high accuracy in predicting students' scores. One of the characteristics of audio is that it can be adopted as additional information, as it is closely associated with text, and the characteristics of changes in sound pulses can be used in the fine classification of emotions. For example, the influence of musical genres on children's emotional intelligence is one of the topics of impact. Qian et al. [93] applied a neural network based on BERT to extract features and analyze emotions in such a way that it was shown to work effectively in the accuracy of musical genre classification tasks based on children's emotions. However, it is difficult for these methods to effectively capture the actual information of the different modalities. Therefore, for the task of the influence of music genres on children's emotions, the BERT transformer was proposed to extract audio-video features and effectively improve the accuracy of sentiment classification tasks.

For sentiment analysis Ding [96] uses students' auditory input, facial expression, and textual data to propose a cross-modal Transformer algorithm to improve information processing. The work is conceived as an innovative idea by using a visual Transformer to achieve accurate and efficient learner emotion analysis in an online teaching context. Achieving an accuracy of 84%. Emotion analysis in this context is crucial for understanding and improving the learning experience and in the evaluation of student engagement in educational courses.

A recent phenomenon is ChatGPT, a conversational artificial intelligence interface chatbot developed by OpenAI. It is being considered as one of the most advanced artificial intelligence applications. ChatGPT is a revolutionary tool that answers questions about almost anything available in the digital environment and can help the state create and implement an impartial and fair curriculum. If properly implemented, it could serve as a bridge to ease the pressure on a stressed education system. Spontaneous student opinions on a particular topic reveal significant data, such as in the case of university teaching practices, where a linguistic model based on AI Generative Pre-trained Transformer-3 (GPT-3) was applied. Sentiment analysis

showed that students prefer clear teaching practices where ideas and activities are presented unambiguously and based on interaction between teachers and students and among students themselves [92].

A study of 11.830 tweets about the use of the new ChatGPT technology revealed that opinions on the application of ChatGPT in education show that many tweets are positive or neutral, with a small percentage expressing negative sentiments [95]. It is important to mention that this article was initially thought to belong to category B because the data was collected from a social network. However, after further analysis of its content, it was decided to categorize it in section C as a new scenario. However, there are some concerns about the application in educational settings due to issues such as the deception, honesty, and truthfulness of ChatGPT [94].

V. Discussion

The purpose of this study was to conduct a systematic review of the literature on sentiment analysis with Transformers models applied to education to gain a better understanding of its status, how can influence culture, Open Science and the development of artificial intelligence, together with its benefits, challenges, and future work. Four broad research questions were specified in the Introduction section which are now addressed.

RQ1 investigated the current state of the art in sentiment analysis with Transformers applied to education. To answer this question, 41 published research articles were examined. The first finding reveals that most studies focus on the sentiment analysis of data collected from online educational platforms. For example, comments [62], messages left on the forums [71], or MOOC chat messages [79] were analyzed. Another scenario from which data were collected was social networks, especially comments on Twitter [87], Facebook, and YouTube [89]. In addition, comments from forums, such as Reddit [82]. Finally, data were taken from environments other than those mentioned above, such as comments left in university departments [100] or surveys given to students to determine their opinions [97]. Therefore, the articles were classified into three categories. Category A corresponds to articles that used data collected from online educational platforms. This category represents 46% of the reviewed studies and deals with sentiment analysis applied in virtual educational environments to improve teaching and learning processes. Category B corresponds to articles that used data collected from social networks with 32% of the studies and reflects sentiment analysis of opinions that students leave on social networks. Finally, category C, with 22% of the studies, groups studies that use data that has been taken from surveys, or opinions of new trends in sentiment analysis with audio [101] and new technologies such as ChatGPT [95].

RQ2 explored the advantages of employing sentiment analysis using transformers in the realm of education. Initially, the main findings from scrutinized research studies underscored the superior performance of transformer models compared to other categories of machine learning or deep learning technologies, achieving a sentiment prediction accuracy of over 90% [99]. The direct merit lies in the ability of the models to enhance the likelihood of accurate sentiment predictions. Upon closer examination of the literature, one notable advantage emerged: the objectivity of the data used for sentiment analysis because there are two main approaches to obtaining student feedback: the direct approach and the indirect approach. In the direct approach, opinions are collected through the distribution of questionnaires and the subsequent collection of responses [97]. However, this method has limitations as it does not reveal the true experience of students and there is a possibility of bias in the collection and evaluation of questionnaires. To overcome these limitations, an indirect approach can be adopted, where social media posts serve as a source for

collecting students' opinions, as students are active on social media and use social media to express their opinions through posts. This objectivity is derived from the source platforms, where students freely express their opinions without external pressures, allowing genuine feelings, perceptions, and opinions to be identified [80]. Furthermore, the investigations revealed multiple benefits derived from the application of sentiment analysis in educational settings. Notably, the methodology facilitates the detection of students' needs, thereby contributing to the enhancement of course quality, refinement of the teaching system, and diversification of instructional materials within online education environments. Consequently, this contributes to heightened satisfaction among students engaging in virtual courses. Indeed, certain studies even delved into the examination of students' phobias, shedding light on issues that warrant attention. In summary, the overarching goal of conducting sentiment analysis is to leverage students' feedback for the continual improvement of educational quality.

RQ3 covered the key challenges in applying sentiment analysis with Transformers in education. There are students with different cultural backgrounds, and identifying online learning environments that respect the particularities of international students is a great challenge [102]. Language is one of the necessary elements for sentiment analysis that reflects culture [103]. First, research on sentiment analysis in online learning environments has predominantly focused on the English language, as it is a language with many available resources, such as reference datasets, annotated corpora, and sentiment lexicons [104], [105]. However, Deriu et al. [106] reported that sentiment analysis methods developed for single-language texts could not be replicated for novel or multilingual texts. In addition, the English language and its common terms and understanding are causing problems for Open Science. For this reason, studies on sentiment analysis in educational settings have also focused on languages other than English. Munshi et al. [97] propose the AraBERT multilingual model based on Arabic, a morphologically rich language with multiple dialects. Similarly, data collected from online educational environments can be multilingual and multicultural, such as those of the Chinese language [70] or Filipino lexicons [89].

Another challenge reflected in the educational context is the concern regarding the use of new artificial intelligence technologies, where sentiment analysis may encounter negative views regarding ethics and dishonesty. Javaid et al. [94] argue that the use of ChatGPT and other linguistic models raises crucial ethical questions about their effects on society. However, there are areas of research to apply new technologies so that students can easily understand and communicate in other languages [94]. At present, the results confirm that ChatGPT is widely used in education [95]. Therefore, new technologies can be used as tools to help students create relevant content on a specific topic. It can also be used to provide students with feedback to help them improve their knowledge based on their moods. This could help ensure that students receive the right amount of challenge and material that is interesting and relevant.

Regarding the Transformer models applied in the research field, the challenge lies in the quality of the training dataset, as the aim is to avoid biased data due to different linguistic and cultural contexts [80]. Therefore, the suggested strategy is to improve the pre-processing stage for correct sentiment analysis. Another aspect is the large number of parameters, which leads to a time-consuming training process [69]. As an alternative, model compression was proposed as far as possible to reduce this time. However, the training process results in high computational costs.

RQ4 addressed the identification of future areas of research. At the end of this systematic literature review, it was determined that there are few studies on sentiment analysis related to culture. The method

applied to analyze learner comments left on MOOCs is based on applying the lexicon in a single language [65]. In the work of Marfani et al. [67] about 10.000 reviews from the Coursera platform were analyzed, and data processing was performed in a single language without considering cultural aspects. Considering these works, there is a research space for sentiment analysis with artificial intelligence that includes the opinions of online learning environments of learners from different cultures and nationalities.

Furthermore, it is recommended to give students control over their learning processes by offering multiple cultural options in MOOCs [107]. This recommendation is related to UNESCO [108] recommendation on Open Science by stating that multilingual scientific knowledge should be openly available, accessible, and reusable. Future research should focus on understanding learners' perceptions of open access to multilingual resources in online or distance learning environments. Understanding cultural differences can improve teaching and learning processes to provide quality and culturally sensitive education [102].

VI. Conclusions, Limitations, and Future Works

In recent years, sentiment analysis has been a breakthrough in the field of natural language processing. This paper presents a literature review on sentiment analysis research utilizing Transformers deep learning models. The scope of the research was limited to the educational field and the 6W systematic review method was used to analyze 41 articles from well-known digital databases such as Scopus, ScienceDirect, IEEE Xplore, SpringerLink, Taylor & Francis, and MDPI. The results explain the current state of knowledge on sentiment analysis with Transformers in education and identify the benefits and challenges of use. Furthermore, future areas of research for this modern AI technology and findings in terms of its implications were identified.

The present study achieved significant results in sentiment analysis within the education domain using Transformer models. However, a notable limitation was the difficulty in accessing and consulting relevant databases to filter studies specifically related to the topic. This challenge hindered the refinement of search processes and the identification of precise findings pertinent to this research.

This study has identified the advantages and challenges of using sentiment analysis in educational settings with Transformers. Therefore, future research directions in sentiment analysis applied to the sustainability of education could focus on refining specialized models, integrating multimodal data, assessing the long-term impact of Open Science initiatives, developing specific metrics, actively involving the educational community, creating interactive tools, applying to distance learning environments, and emphasizing inclusion and cultural diversity. In essence, the integration of sentiment analysis with AI technologies heralds a paradigm shift in educational research and practice, empowering stakeholders to forge a more responsive, empathetic, and student-centric educational landscape. As stride steadfastly into the digital age, the synergy between AI and education promises to redefine the contours of teaching and learning, ushering in an era of unprecedented innovation and transformation.

Acknowledgement

References

[1] T. Nasukawa and J. Yi, "Sentiment analysis: capturing favorability using natural language processing," in *Proceedings of the 2nd International Conference on Knowledge Capture*, in K-CAP '03. New York, NY, USA: Association for Computing Machinery, 2003, pp. 70–77. doi: 10.1145/945645.945658.

[2] Z. Nanli, Z. Ping, L. Weiguo, and C. Meng, "Sentiment analysis: A literature review," in *2012 International Symposium on Management of Technology (ISMOT)*, 2012, pp. 572–576. doi: 10.1109/ISMOT.2012.6679538.

[3] H. Ruiz Martín, *Aprendiendo a aprender: Mejora tu capacidad de aprender descubriendo cómo aprende tu cerebro*, Vergara. 2020.

[4] T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, "Sentiment analysis and opinion mining on educational data: A survey," *Natural Language Processing Journal*, vol. 2, p. 100003, 2023, doi: https://doi.org/10.1016/j.nlp.2022.100003.

[5] Z. Han, J. Wu, C. Huang, Q. Huang, and M. Zhao, "A review on sentiment discovery and analysis of educational big-data," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, Jul. 11, 2019. doi: 10.1002/widm.1328.

[6] C. Yan, J. Liu, W. Liu, and X. Liu, "Sentiment Analysis and Topic Mining Using a Novel Deep Attention-Based Parallel Dual-Channel Model for Online Course Reviews," *Cognitive Computation*, vol. 15, no. 1, pp. 304–322, 2023. doi: 10.1007/s12559-022-10083-7.

[7] B. Du, "Research on the factors influencing the learner satisfaction of MOOCs," *Education and Information Technologies*, vol. 28, no. 2, pp. 1935–1955, 2023. doi: 10.1007/s10639-022-11269-0.

[8] J. Vansina, "Oral Tradition as History," *University of Wisconsin Press*, pp. 0–258, 1985.

[9] J.-B. Michel *et al.*, "Quantitative Analysis of Culture Using Millions of Digitized Books," *Science*, vol. 331, no. 6014, pp. 176–182, Jan. 14, 2011. doi: 10.1126/science.1199644.

[10] Costa Albert *et al.*, "Your Morals Depend on Language," *PLoS ONE*, vol. 9, no. 4, Public Library of Science, pp. 1–7, Aug. 2014. doi: 10.1371/journal.pone.0094842.

[11] R. J. Lennox, D. Veríssimo, W. M. Twardek, C. R. Davis, and I. Jarić, "Sentiment analysis as a measure of conservation culture in scientific literature," *Conservation Biology*, vol. 34, no. 2, pp. 462–471, 2020, doi: https://doi.org/10.1111/cobi.13404.

[12] Ravi. Vatrapu, "Cultural Considerations in Computer Supported Collaborative Learning," *Research and Practice in Technology Enhanced Learning*, vol. 03, no. 02, pp. 159–201, Jul. 2008. doi: 10.1142/S1793206808000501.

[13] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017.

[14] M. Mansoori, H. Maliwal, S. Kotian, H. Kenkre, I. Saha, and P. Mishra, "A Systematic Survey on Computational agents for Mental Health Aid," in *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, 2022, pp. 1–7. doi: 10.1109/I2CT54291.2022.9824269.

[15] E. Oghu, E. Ogbuju, T. Abiodun, and F. Oladipo, "A Review of Sentiment Analysis Approaches for Quality Assurance in Teaching and Learning," *Bulletin of Social Informatics Theory and Application*, vol. 6, no. 2, pp. 177–188, Jan. 2023, doi: https://doi.org/10.31763/businta.v6i2.581.

[16] T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, "Sentiment analysis and opinion mining on educational data: A survey," *Natural Language Processing Journal*, vol. 2, p. 100003, 2023, doi: https://doi.org/10.1016/j.nlp.2022.100003.

[17] S. Saeedi, "Socially Aware Natural Language Processing with Commonsense Reasoning and Fairness in Intelligent Systems," Western Michigan University, 2023.

[18] P. P. Dakle *et al.*, "Ner4Opt: Named Entity Recognition for Optimization Modelling from Natural Language," in *Integration of Constraint Programming, Artificial Intelligence, and Operations Research: 20th International Conference*, Berlin, Heidelberg: Springer-Verlag, 2023, pp. 299–319. doi: 10.1007/978-3-031-33271-5_20.

[19] T. Firoozi, H. Mohammadi, and M. J. Gierl, "Using Active Learning Methods to Strategically Select Essays for Automated Scoring," *Educational Measurement: Issues and Practice*, vol. 42, no. 1, pp. 34–43, Dec. 2022, doi: 10.1111/emip.12537.

[20] S. Prabhu, K. Akhila, and S. S, "A Hybrid Approach Towards Automated Essay Evaluation based on Bert and Feature Engineering," in *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, 2022, pp. 1–4. doi: 10.1109/I2CT54291.2022.9824999.

[21] M. A. Sayeed and D. Gupta, "Automate Descriptive Answer Grading using Reference based Models," in *2022 OITS International Conference on Information Technology (OCIT)*, 2022, pp. 262–267. doi: 10.1109/OCIT56763.2022.00057.

[22] E. Andersson, "Methods for increasing cohesion in automatically extracted summaries of Swedish news articles : Using and extending multilingual sentence transformers in the data-processing stage of training BERT models for extractive text summarization," Linköping University, Department of Computer and Information Science, 2022.

[23] L. Mo *et al.*, "Incorporating Sentimental Trend into Gated Mechanism Based Transformer Network for Story Ending Generation," *Neurocomputing*, vol. 453, Jan. 2021, doi: 10.1016/j.neucom.2021.01.040.

[24] A. Singh and G. Josan, "Paraphrase Generation: A Review from RNN to Transformer based Approaches," *International Journal of Next-Generation Computing*, Apr. 2022, doi: 10.47164/ijngc.v13i1.377.

[25] I. Cid Rico and J. Pascual Espada, "Pretrained transformers models for extracting keywords in educational texts." Accessed: Aug. 25, 2024. [Online]. Available: https://ssrn.com/abstract=4442925

[26] M. U. Demirezen, O. Yilmaz, and E. Ince, "New models developed for detection of misconceptions in physics with artificial intelligence," *Neural Computing and Applications*, vol. 35, no. 12, pp. 9225–9251, 2023. doi: 10.1007/s00521-023-08414-2.

[27] B. Ahmed, M. Saad, and E. A. Refaee, "QQATeam at Qur'an QA 2022: Fine-Tunning Arabic QA Models for Qur'an QA Task," in *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, Marseille, France: European Language Resources Association, Jun. 2022, pp. 130–135. Accessed: Aug. 25, 2024. [Online]. Available: https://aclanthology.org/2022.osact-1.16

[28] A. Coenen *et al.*, "Visualizing and measuring the geometry of BERT," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2019.

[29] J. Chun, "SentimentArcs: A Novel Method for Self-Supervised Sentiment Analysis of Time Series Shows SOTA Transformers Can Struggle Finding Narrative Arcs," *ArXiv*, 2021. Accessed: Feb. 03, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2110.09454

[30] A. Nagarajan, S. Sen, J. R. Stevens, and A. Raghunathan, "Specialized Transformers: Faster, Smaller and more Accurate NLP Models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=aUoV6qhY_e

[31] P. Bai, Y. Xia, and Y. Xia, "Fusing Knowledge and Aspect Sentiment for Explainable Recommendation," *IEEE Access*, vol. 8, pp. 137150–137160, 2020, doi: 10.1109/ACCESS.2020.3012347.

[32] Q. Dong, T. Sun, Y. Xu, X. Xu, M. Zhong, and K. Yan, "Network Public Opinion Sentiment Analysis based on Bert Model," in *2022 IEEE 10th International Conference on Information, Communication and Networks (ICICN)*, 2022, pp. 662–666. doi: 10.1109/ICICN56848.2022.10006589.

[33] P. Ekman *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion," *Journal of Personality and Social Psychology*, vol. 53, no. 4, pp. 712–717, 1987. doi: 10.1037/0022-3514.53.4.712.

[34] C. E. Izard, D. Z. Libero, P. Putnam, and O. M. Haynes, "Stability of emotion experiences and their relations to traits of personality.," *Journal of Personality and Social Psychology*, vol. 64, no. 5, pp. 847–860, 1993. doi: 10.1037/0022-3514.64.5.847.

[35] O. Bruna, H. Avetisyan, and J. Holub, "Emotion models for textual emotion classification," *Journal of Physics: Conference Series*, vol. 772, p. 012063, Nov. 2016. doi: 10.1088/1742-6596/772/1/012063.

[36] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Association for Computational Linguistics (ACL), 2019, pp. 4171–4186.

[37] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," in *International Conference on Learning Representations*, Jul. 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[38] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," in *International Conference on Learning Representations*, Sep. 2019. doi: 10.48550/arxiv.1909.11942.

[39] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Jun. 2019. [Online]. Available: http://arxiv.org/abs/1906.08237

[40] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *ArXiv*, Oct. 02, 2019. Accessed: Jan. 20, 2024. [Online]. Available: http://arxiv.org/abs/1910.01108

[41] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The Efficient Transformer," in *International Conference on Learning Representations*, 2020. Accessed: Aug. 25, 2024. [Online]. Available: https://arxiv.org/abs/2001.04451

[42] P. Budzianowski and I. Vulić, "Hello, It's GPT-2 – How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems," in *Proceedings of the 3rd Workshop on Neural Generation and Translation*, 2019, pp. 15–22. doi: 10.18653/v1/D19-5602.

[43] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/145 7c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[44] G. Yenduri, B. R. Rajakumar, K. Praghash, and D. Binu, "Heuristic-Assisted BERT for Twitter Sentiment Analysis," in *International Journal of Computational Intelligence and Applications*, 2021. doi: 10.1142/S1469026821500152.

[45] X. Gong, W. Ying, S. Zhong, and S. Gong, "Text Sentiment Analysis Based on Transformer and Augmentation," *Frontiers in Psychology*, vol. 13, Sep. 2022. doi: 10.3389/fpsyg.2022.906061.

[46] N. Pahari and K. Shimada, "Multi-Task Learning Using BERT With Soft Parameter Sharing Between Layers," in *International Conference on Soft Computing and Intelligent Systems* , 2022, pp. 1–6. doi: 10.1109/SCISISIS55246.2022.10001943.

[47] S. Zanwar, D. Wiechmann, Y. Qiao, and E. Kerz, "Improving the Generalizability of Text-Based Emotion Detection by Leveraging Transformers with Psycholinguistic Features," in *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science*, 2022, pp. 1–13. doi: 10.18653/v1/2022.nlpcss-1.1.

[48] J. Shen, X. Liao, and Z. Tao, "Sentence-level sentiment analysis via BERT and BiGRU," in *2019 International Conference on Image and Video Processing, and Artificial Intelligence*, R. Su, Ed., in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 11321. Nov. 2019, p. 113212S. doi: 10.1117/12.2550215.

[49] S. Mol and P. S. Sreeja, "FNDNLSTM: Fake News Detection on Social Media Using Deep Learning (DL) Techniques," in *New Opportunities for Sentiment Analysis and Information Processing*, 2021, pp. 218–232. doi: 10.4018/978-1-7998-8061-5.ch012.

[50] P. N. Begen and L. Vidiasova, "Development of an Algorithm for Fixing the Citizens' Assessments of Digital Transformation Processes Based on Text Analysis," in *Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance*, in ICEGOV '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 585–587. doi: 10.1145/3560107.3560203.

[51] S. Li, W. Deng, and J. Hu, "Momentum Distillation Improves Multimodal Sentiment Analysis," in *Pattern Recognition and Computer Vision*, S. Yu, Z. Zhang, P. C. Yuen, J. Han, T. Tan, Y. Guo, J. Lai, and J. Zhang, Eds., Cham: Springer International Publishing, 2022, pp. 423–435.

[52] L. Bacco, A. Cimino, F. Dell'Orletta, and M. Merone, "Extractive Summarization for Explainable Sentiment Analysis using Transformers," in *Sixth International Workshop on Explainable Sentiment Mining and Emotion detection*, 2021. [Online]. Available: https://openreview.net/forum?id=xB1deFXLaF9

[53] X. Zhu, Y. Zhu, L. Zhang, and Y. Chen, "A BERT-based multi-semantic learning model with aspect-aware enhancement for aspect polarity classification," *Applied Intelligence*, vol. 53, no. 4, pp. 4609–4623, Feb. 2023, doi: 10.1007/s10489-022-03702-1.

[54] Z. Wu and D. C. Ong, "Context-Guided BERT for Targeted Aspect-Based Sentiment Analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, May 2021, pp. 14094–14102. doi: 10.1609/aaai.v35i16.17659.

[55] J. Yang *et al.*, "Multi-Applicable Text Classification Based on Deep Neural Network," *International Journal of Sensor Networks*, vol. 40, no. 4, Inderscience Publishers, Geneva 15, CHE, pp. 277–286, Jan. 2022. doi: 10.1504/ijsnet.2022.127841.

[56] Z. Yuan, W. Li, H. Xu, and W. Yu, "Transformer-Based Feature Reconstruction Network for Robust Multimodal Sentiment Analysis," in *Proceedings of the 29th ACM International Conference on Multimedia*, in MM '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 4400–4407. doi: 10.1145/3474085.3475585.

[57] P. Howard *et al.*, "Cross-Domain Aspect Extraction using Transformers Augmented with Knowledge Graphs," in *International Conference on Information and Knowledge Management, Proceedings*, Association for Computing Machinery, Oct. 2022, pp. 780–790. doi: 10.1145/3511808.3557275.

[58] C. Zhao, S. Wang, D. Li, X. Liu, X. Yang, and J. Liu, "Cross-Domain Sentiment Classification via Parameter Transferring and Attention Sharing Mechanism," *Information Sciences*, vol. 578, no. C, Elsevier Science Inc., USA, pp. 281–296, Nov. 2021. doi: 10.1016/j.ins.2021.07.001.

[59] J. Paul and A. R. Criado, "The art of writing literature review: What do we know and what do we need to know?," *International Business Review*, vol. 29, no. 4, p. 101717, 2020, doi: https://doi.org/10.1016/j.ibusrev.2020.101717.

[60] J. L. Callahan, "Writing Literature Reviews: A Reprise and Update," *Human Resource Development Review*, vol. 13, no. 3, pp. 271–275, 2014, doi: 10.1177/1534484314536705.

[61] M. Rodenes Adam, R. Salvador Vallés, G. I. Moncaleano Rodríguez, "E-learning: características y evaluación," *Ensayos de Economía*, vol. 23, no. 43, pp. 143-160, 2013

[62] Y. Zhou and M. Li, "Online Course Quality Evaluation Based on BERT," in *2020 International Conference on Communications, Information System and Computer Engineering (CISCE)*, 2020, pp. 255–258. doi: 10.1109/CISCE50729.2020.00057.

[63] C. Yan, J. Liu, W. Liu, and X. Liu, "Sentiment Analysis and Topic Mining Using a Novel Deep Attention-Based Parallel Dual-Channel Model for Online Course Reviews," *Cognitive Computation*, vol. 15, no. 1, pp. 304–322, 2023, doi: 10.1007/s12559-022-10083-7.

[64] D. Jatain, V. Singh, and N. Dahiya, "An exploration of the causal factors making an online course content popular & engaging," *International Journal of Information Management Data Insights*, vol. 3, no. 2, p. 100194, 2023, doi: https://doi.org/10.1016/j.jjimei.2023.100194.

[65] X. Li, H. Zhang, Y. Ouyang, X. Zhang, and W. Rong, "A Shallow BERT-CNN Model for Sentiment Analysis on MOOCs Comments," in *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*, 2019, pp. 1–6. doi: 10.1109/TALE48000.2019.9225993.

[66] A. Koufakou, "Deep learning for opinion mining and topic classification of course reviews," *Education and Information Technologies*, 2023. doi: 10.1007/s10639-023-11736-2.

[67] H. Marfani, S. Hina, and H. Tabassum, "Analysis of Learners' Sentiments on MOOC Forums using Natural Language Processing Techniques," in *2022 3rd International Conference on Innovations in Computer Science & Software Engineering (ICONICS)*, 2022, pp. 1–8. doi: 10.1109/ICONICS56716.2022.10100401.

[68] X. Pan, B. Hu, Z. Zhou, and X. Feng, "Are students happier the more they learn? – Research on the influence of course progress on academic emotion in online learning," *Interactive Learning Environments*, pp. 1–21, 2022, doi: 10.1080/10494820.2022.2052110.

[69] T. Liu, W. Hu, H. Guo, and Y. Li, "Sentiment Analysis of MOOC Reviews Based On Capsule Network," in *2021 4th International Conference on Intelligent Autonomous Systems (ICoIAS)*, 2021, pp. 222–227. doi: 10.1109/ICoIAS53694.2021.00047.

[70] L. Min, X. Miao, P. Bi, and F. He, "A Small amount of Labeled Data Chinese Online Course Review Target Extraction via ALBERT-IDCNN-CRF Model," *Journal of Physics: Conference Series*, vol. 1651, no. 1, IOP Publishing, p. 012049, 2020. doi: 10.1088/1742-6596/1651/1/012049.

[71] N. A. Khodeir, "Bi-GRU Urgent Classification for MOOC Discussion Forums Based on BERT," *IEEE Access*, vol. 9, pp. 58243–58255, 2021, doi: 10.1109/ACCESS.2021.3072734.

[72] S. Liu, S. Liu, Z. Liu, X. Peng, and Z. Yang, "Automated detection of emotional and cognitive engagement in MOOC discussions to predict learning achievement," *Computers & Education*, vol. 181, p. 104461, 2022. doi: https://doi.org/10.1016/j.compedu.2022.104461.

[73] Z. Liu, X. Kong, H. Chen, S. Liu, and Z. Yang, "MOOC-BERT: Automatically Identifying Learner Cognitive Presence From MOOC Discussion Data," *IEEE Transactions on Learning Technologies*, vol. 16, no. 4, pp. 528–542, 2023, doi: 10.1109/TLT.2023.3240715.

[74] Y. Qu, F. Li, L. Li, X. Dou, and H. Wang, "Can We Predict Student Performance Based on Tabular and Textual Data?," *IEEE Access*, vol. 10, pp. 86008–86019, 2022, doi: 10.1109/ACCESS.2022.3198682.

[75] N. Alkaabi, N. Zaki, H. Ismail, and M. Khan, "Detecting Emotions behind the Screen," *AI*, vol. 3, no. 4, pp. 948–960, 2022, doi: 10.3390/ai3040056.

[76] R. Sanchis-Font, M. J. Castro-Bleda, J.-Á. González, F. Pla, and L.-F. Hurtado, "Cross-Domain Polarity Models to Evaluate User eXperience in E-learning," *Neural Processing Letters*, vol. 53, no. 5, pp. 3199–3215, 2021. doi: 10.1007/s11063-020-10260-5.

[77] Y. Y. Dyulicheva, "Learning Analytics in MOOCs as an Instrument for Measuring Math Anxiety," *Voprosy Obrazovaniya Educational Studies Moscow*, no. 4, pp. 132–147, 2021, doi: 10.17323/1814-9545-2021-4-243-265.

[78] D. and T. D. and C. I. Nikolovski Vlatko and Kitanovski, "Case Study: Predicting Students Objectivity in Self-evaluation Responses Using Bert Single-Label and Multi-Label Fine-Tuned Deep-Learning Models," in *ICT Innovations 2020. Machine Learning and Applications*, I. Dimitrova Vesna and Dimitrovski, Ed., Cham: Springer International Publishing, 2020, pp. 98–110.

[79] B. Su and J. Peng, "Sentiment Analysis of Comment Texts on Online Courses Based on Hierarchical Attention Mechanism," *Applied Sciences*, vol. 13, no. 7, 2023, doi: 10.3390/app13074204.

[80] S. B. A. Prasad and R. P. K. Nakka, "Supervised Sentiment Analysis of Indirect Qualitative Student Feedback for Unbiased Opinion Mining," *Engineering Proceedings*, vol. 59, no. 1, 2023, doi: 10.3390/engproc2023059015.

[81] M. Alzaid and F. Fkih, "Sentiment Analysis of Students’ Feedback on E-Learning Using a Hybrid Fuzzy Model," in *Applied Sciences*, 2023. doi: 10.3390/app132312956.

[82] M. Y. Hameed, L. Al-Hindi, S. Ali, H. K. Jensen, and C. C. Shoults, "Broadening the Understanding of Medical Students' Discussion of Radiology Online: A Social Listening Study of Reddit," *Current problems in diagnostic radiology*, vol. 52, no. 5, p. 377—382, 2023. doi: 10.1067/j.cpradiol.2023.04.003.

[83] R. Jyothsna, V. Rohini, and J. Paulose, "Sentiment Analysis of Stress Among the Students Amidst the Covid Pandemic Using Global Tweets," in *Ambient Intelligence in Health Care*, T. Swarnkar, S. Patnaik, P. Mitra, S. Misra, and M. Mishra, Eds., Singapore: Springer Nature Singapore, 2023, pp. 317–324.

[84] H. and Ö. O. and G. A. C. and K. B. and Ş. Ö. and A. F. P. Yldrm Elif and Yazgan, "Sentiment Analysis of Tweets on Online Education during COVID-19," in *Artificial Intelligence Applications and Innovations*, L. and M. J. and D. M. Maglogiannis Ilias and Iliadis, Ed., Cham: Springer Nature Switzerland, 2023, pp. 240–251.

[85] M. Mujahid *et al.*, "Sentiment Analysis and Topic Modeling on Tweets about Online Education during COVID-19," *Applied Sciences*, vol. 11, no. 18, 2021, doi: 10.3390/app11188438.

[86] M. Zhou and H. Mou, "Tracking public opinion about online education over COVID-19 in China," *Educational Technology Research and Development*, vol. 70, no. 3, pp. 1083–1104, Jun. 2022. doi: 10.1007/s11423-022-10080-5.

[87] V. Duong, J. Luo, P. Pham, T. Yang, and Y. Wang, "The Ivory Tower Lost: How College Students Respond Differently than the General Public to the COVID-19 Pandemic," in *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2020, pp. 126–130. doi: 10.1109/ASONAM49781.2020.9381379.

[88] Y. Ren and X. Tan, "Research on the Method of Identifying Students' Online Emotion Based on ALBERT," in *2021 International Conference on Intelligent Computing, Automation and Applications (ICAA)*, 2021, pp. 646–650. doi: 10.1109/ICAA53760.2021.00118.

[89] L. L. Maceda, A. A. Satuito, and M. B. Abisado, "Sentiment Analysis of Code-mixed Social Media Data on Philippine UAQTE using Fine-tuned mBERT Model," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 7, 2023, doi: 10.14569/IJACSA.2023.0140777.

[90] A. S. Mondal, Y. Zhu, K. K. Bhagat, and N. Giacaman, "Analysing user reviews of interactive educational apps: a sentiment analysis approach," *Interactive Learning Environments*, vol. 32, no. 1, pp. 355-372, 2022, doi: 10.1080/10494820.2022.2086578.

[91] S. Fouad and E. Alkooheji, "Sentiment Analysis for Women in STEM using Twitter and Transfer Learning Models," in *Proceedings - 17th IEEE International Conference on Semantic Computing, ICSC 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 227–234. doi: 10.1109/ICSC56153.2023.00045.

[92] C. Álvarez-Álvarez and S. Falcon, "Students' preferences with university teaching practices: analysis of testimonials with artificial intelligence," *Educational technology research and development*, 2023, doi: 10.1007/s11423-023-10239-8.

[93] Q. Qian and X. Chen, "A Multi-Modal Convolutional Neural Network Model for Intelligent Analysis of the Influence of Music Genres on Children's Emotions," *Computational Intelligence and Neuroscience*, vol. 2022, Jul. 2022. doi: 10.1155/2022/4957085.

[94] M. Javaid, A. Haleem, R. P. Singh, S. Khan, and I. H. Khan, "Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, vol. 3, no. 2, p. 100115, 2023, doi: https://doi.org/10.1016/j.tbench.2023.100115.

[95] M. Tubishat, F. Al-Obeidat, and A. Shuhaiber, "Sentiment Analysis of Using ChatGPT in Education," in *2023 International Conference on Smart Applications, Communications and Networking (SmartNets)*, 2023, pp. 1–7. doi: 10.1109/SmartNets58706.2023.10215977.

[96] L. Ding, "Online teaching emotion analysis based on GRU and nonlinear transformer algorithm," *PeerJ Computer Science*, vol. 9, PeerJ Inc., 2023. doi: 10.7717/peerj-cs.1696.

[97] A. Munshi *et al.*, "Arabic Sentiment Analysis for Student Evaluation using Machine Learning and the AraBERT Transformer," *Engineering, Technology and Applied Science Research*, vol. 13, no. 5, pp. 11945 – 11952, 2023, doi: 10.48084/etasr.6347.

[98] C. Zhu, "Analysis of the Sentiment in the Evaluation Texts of University Students by Means of the Concept of Flexible Management," *International Journal of Emerging Technologies in Learning*, vol. 18, no. 18, pp. 261–276, 2023, doi: 10.3991/ijet.v18i18.43505.

[99] C.-Y. Chou and T.-Y. Chuang, "Aspect and Sentiment Classification Mechanisms of Student After-Class Self-Evaluated Comments: Investigation on Nonsense Data, Feature Extraction, and Classification Models," *Engineering Proceedings*, vol. 38, no. 1, 2023, doi: 10.3390/engproc2023038043.

[100] H. Setiawan, C. Fatichah, and A. Saikhu, "Multilabel Classification of Student Feedback Data Using BERT and Machine Learning Methods," in *2023 14th International Conference on Information & Communication Technology and System (ICTS)*, 2023, pp. 147–152. doi: 10.1109/ICTS58770.2023.10330849.

[101] D. Chowdary Attota and N. Dehbozorgi, "Towards Application of Speech Analysis in Predicting Learners' Performance," in *2022 IEEE Frontiers in Education Conference (FIE)*, 2022, pp. 1–5. doi: 10.1109/FIE56618.2022.9962701.

[102] P. Gómez-Rey, E. Barbera, and F. Fernández-Navarro, "The impact of cultural dimensions on online learning," *Educational Technology & Society*, vol. 19, pp. 225–238, Jan. 2016.

[103] W. Jiang, "The relationship between culture and language," *ELT journal*, vol. 54, no. 4, pp. 328–334, 2000.

[104] M. M. Agüero-Torales, J. I. Abreu Salas, and A. G. López-Herrera, "Deep learning and multilingual sentiment analysis on social media data: An overview," *Applied Soft Computing*, vol. 107, p. 107373, 2021. doi: https://doi.org/10.1016/j.asoc.2021.107373.

[105] K. R. Mabokela, T. Celik, and M. Raborife, "Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape," *IEEE Access*, vol. 11, pp. 15996–16020, 2023, doi: 10.1109/ACCESS.2022.3224136.

[106] J. Deriu *et al.*, "Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification," in *Proceedings of the 26th International Conference on World Wide*, 2017. [Online]. Available: https://arxiv.org/abs/1703.02504

[107] R. Shahini, H. Davis, and K. Borthwick, "Design recommendations to address cultural issues in multicultural MOOCs: a systematic literature review," *New educational landscapes: innovative perspectives in language learning and technology*, pp. 55–66, 2019.

[108] UNESCO, "UNESCO Recommendation on Open Science." Accessed: Aug. 24, 2024. [Online]. Available: https://www.unesco.org/en/open-science/about

Anabel Pilicita

Anabel Pilicita obtained a Master's degree in Network Engineering and Telematic Services at the Universidad Politécnica de Madrid (UPM) in 2016, where he is currently pursuing a Ph.D. in Telematic Services Engineering. His research interests include the application of natural language processing and new artificial intelligence models.

Enrique Barra

Enrique Barra received the Ph.D. degree in telematics engineering with a minor in multimedia and technology enhanced learning from the Universidad Politécnica de Madrid (UPM). He has participated in many European projects, such as GLOBAL, FIWARE, and C@R. He is currently involved in several projects contributing to the generation and distribution of educational content in TEL environments. His research interests include videoconferencing, games in education, and social networks in education.

Youth Expectations and Perceptions of Generative Artificial Intelligence in Higher Education

Andrea E. Cotino-Arbelo, Carina S. González-González, Jezabel Molina-Gil *

Women's Research Institute, University of La Laguna, La Laguna (Spain)

* Corresponding author: acotinoa@ull.edu.es (A. E. Cotino Arbelo), cjgonza@ull.edu.es (C. S. González-González), jmmolina@ull.edu.es (J. Molina Gil).

Abstract

Artificial Intelligence (AI) is not a recent innovation, what's new is how accessible its features have become across multiple devices, apps, and services. Sensationalistic news can distort public perception by exaggerating AI's capabilities and risks. This leads to misconceptions and unrealistic expectations, causing misunderstandings about the true nature and limitation of these tools. Such distortions can undermine trust and hinder the effective adoption and integration of AI into society. This study aims to address this issue by exploring the expectations and perceptions of young individuals regarding Generative Artificial Intelligence (GAI) tools. It explores their understanding of GAI and related devices, such as virtual assistants, chatbots, and social robots, which can incorporate GAI. A total of $N=100$ university students engaged in this study by completing a digital questionnaire distributed through the virtual campus of the University of La Laguna. The quantitative analysis uncovered a significant gap in participants' understanding of GAI terminology and its underlying mechanisms. Additionally, it shed light on a noteworthy gender-based discrepancy in the expressed concerns. Participants commonly recognized their ability to communicate effectively with GAI, asserting that such interactions enhance their emotional well-being. Notably, virtual assistants and chatbots were perceived as more valuable tools compared to social robots within the educational realm.

Keywords

I. Introduction

SCIENCE unfolds a realm of opportunities to delve into and understand the foundations of Generative Artificial Intelligence (GAI onwards) tools across diverse domains [1]. While it is true that these tools provide the opportunity to address complex challenges without restrictions in scope or strict theoretical knowledge, we encounter the presence of journalists seeking to capture lectors' attention with sensationalistic news about the impact of GAI tools on society [1]-[3]. Journalists present these tools as a fascinating technology that, at best, promises to improve our quality of life [1] [4]-[7]. This approach hinders the user's real understanding of the true nature of GAI tools, creating a gap between expectations and reality.

The lack of real contextualization may contribute to a distorted perception of GAI tools' actual capabilities and limitations. Furthermore, false promises can undermine people's trust in emerging technology, as the hopes placed in expectations are significantly linked to trust [8]. Promises are meant to foster beliefs in future actions, which can influence the user's perception, emotion, and behavior in multiple ways [9]. While GAI tools offer innovative solutions [3], we must not forget that overlooking their limitations and risks can negatively shape an individual's perception of these tools and, in some cases, even of human capabilities [10].

To understand the emerging technological paradigm, it is crucial to analyze the dynamics of user expectations related to these tools [11], particularly among the younger population. Therefore, the aim of this pilot study is to identify the expectations and perceptions of young individuals regarding GAI tools, as well as the understanding of GAI and devices such as virtual assistants, chatbots, and social robots. What are the expectations and perceptions of the youth regarding GAI? This study aims to address this research question, providing an insight into how young individuals currently interact with these technologies.

The structure of the paper is as follows. In Section II, we provide an overview of the framework. Section III details the research methodology employed in this study. Moving forward, Section IV presents the obtained results. Finally, in Sections V and VI we delve into discussions and present conclusions derived from the analysis.

II. Framework

A. A Brief Introduction to GAI

Artificial Intelligence (AI onwards) is not an innovation in its essence. It was first introduced in 1956 by Marvin Minsky and John McCarthy in the Dartmouth Summer Research Project on AI [12] [13]. Since then, its development has been continuous, and significant advancements have occurred over the decades. However, the newness now lies in the ease with which we can employ its functionalities through multiple devices, applications, and AI-driven services.

Its definition has evolved over time, but currently, AI is defined as the ability of machines to mimic certain functionalities of human intelligence, including capabilities such as perception, learning, reasoning, and problem-solving [14][15]. While we can approach and classify AI in multiple ways, this study will focus on two categories: Narrow or Weak AI and General or Strong AI.

Narrow AI represents a more specific and specialized form of AI, designed to perform specific tasks within a defined scope. This category excels in specific contexts, such as facial recognition, natural language processing, or medical diagnostics. Narrow AI lacks versatility and adaptability. On the other hand, General AI aspires to achieve a level of intelligence comparable to or even exceeding human intelligence, capable of addressing a wide range of tasks and autonomously learning in diverse contexts. This category aims to replicate human intelligence's versatility and adaptability characteristics, enabling machines to perform specific tasks while understanding, reasoning, and problem-solving more broadly. While Narrow AI focuses on specialization, General AI seeks to mimic human intelligence, posing ethical, technical, and philosophical challenges that are constantly evolving as we progress in this field of study [15].

In this constant evolution, Predictive AI and GAI emerge as two new dimensions within the field of AI. Predictive AI focuses on the ability to anticipate future events through analysis of patterns and historical data. However, GAI goes beyond replicating predefined functions, enabling machines to generate original and creative content. This opens up new possibilities in art creation, text composition, and design, raises the potential for innovation, and closes collaboration with human creativity [12] [15].

It is important to note that, in the present study, AI is understood as a field of computer science that deals with creating and developing systems and programs capable of performing tasks that typically require human intelligence. While GAI is considered a specific approach within the broader AI field, focusing on the capacity to autonomously create, produce, or generate content.

B. GAI Shaping Today's World

Many of the currently available and operational AI applications are examples of Narrow AI tools designed to perform specific and limited tasks [15]. However, the rapidly growing number of GAI tools available have expanded into society at an overwhelming pace [16]. and access information. These tools' implementation and use have transformed how we interact with technology, access information, and complete our tasks [17]. Establishing boundaries between the human and the artificial has become a challenge by providing coherent responses, simulating emotions, and even generating creative content [18].

In the healthcare domain, GAI plays a crucial role in the analysis of medical images, disease prediction, and treatment personalization [19]. In the financial sector, GAI has become an invaluable tool for data analysis, decision-making, and precise financial planning. Simultaneously, the use of these tools enables the creation of seamless and personalized experiences, fostering higher consumer loyalty, a positive brand perception, and sustainable growth [20]. In the educational sphere, GAI brings significant benefits by allowing the personalization of educational content and the creation of virtual assistants that facilitate interactive learning [17][21]. Furthermore, in the creative industry, GAI is employed for the creation of artistic content, spanning areas such as scriptwriting, filmmaking, journalism, text generation, as well as the creation of music, images, and animations [22].

Regarding GAI devices, virtual assistants and chatbots have stood out among users because they offer a natural and intuitive way of communication with technology [23]. Virtual assistants, such as Siri, Google Assistant, and Alexa (a few of the most popular ones), have transformed how users perform their daily tasks. These assistants can answer questions, engage in real-time conversations, and execute specific actions [24]. Their ability to understand natural language has resulted in a smoother and more accessible user experience, defining a new interaction process between individuals and intelligent machines [25]. Virtual assistants are understood as a technological device interacted with through voice or text commands. On the other hand, chatbots have also become a valuable tool for users. These programs can engage in textual conversations with users, providing quick and efficient responses to their queries [26]. Chatbots are recognized as a program designed to offer assistance through text, with varying levels of intelligence [27]. Furthermore, social robots incorporating GAI are experiencing a significant growth in adoption across various domains. Their ability to act as assistants or companions redefines the Human-Robot Interaction (HRI onwards) process [28]. Social robots are comprehended as robots designed to interact and communicate with people in a more natural manner, resembling human interaction, though not necessarily humanoid in form [29].

As this technology evolves, a future is envisioned where these devices play increasingly integral roles in our everyday lives.

C. Beyond the Books: GAI Education

The introduction of GAI devices and tools has led to significant transformations in the educational sphere. The GAI's ability to generate original content and adapt to users' specific needs has redefined the approach to educational processes [30]. The possibility of creating personalized materials and developing interactive learning experiences has become a feasible and easily accessible task [21]. However, a lack of understanding of the nature of these tools can lead to a range of ethical issues in academic settings [31].

In the First Draft of the Recommendation on the Ethics of AI [14], ten fundamental principles were established to ensure ethics in the development and application of AI. These principles addressed various concerns, from the unintentional reproduction of biases to issues related to the applicability and transparency of technologies, encompassing safety and protection, privacy, oversight, and human decision-making. Additionally, the importance of awareness and AI literacy was taken into account, emphasizing the need to consider multiple ethical aspects in the design and adoption of these innovative technologies in education [14].

In response to the challenges posed within the educational domain, organizations such as UNESCO have issued reports highlighting key points for the proper implementation of AI tools in preschool, primary, and secondary stages [14]. The report acknowledges that the rapid technological evolution may create knowledge gaps in this regard. This initiative aims to address the potential knowledge gaps and foster an informed and proactive dialogue among professionals, thereby contributing to building a solid foundation for the effective adoption of AI in educational environments.

The presence of AI and GAI in the university setting is self-evident [32]. Although these tools allow students to carry out academic tasks more efficiently, they also face the challenge of potential errors or

failures in content generation. In fact, despite being aware of the lack of reliability in responses, some participants still demonstrate a consistent trust in these tools for specific tasks [33]. Universities now emerge as key environments where GAI can play a significant role in research, creativity, and the training of future professionals. Proactively addressing the integration of GAI in higher education is essential to prepare students for the challenges and opportunities that this technology presents to society. Indeed, those who do not incorporate it into classrooms will face a significant disadvantage in the job market [15].

D. Risks to Mitigate

While GAI can benefit and enhance our lives, it also presents risks and challenges that must be carefully addressed, regardless of the professional sector.

GAI tools and devices often interact with personal information. This implies the possibility of data compromise or improper use, failing to safeguard user privacy and security. Simultaneously, the massive collection of data to improve GAI tools raises ethical questions about how these data are stored, used, and shared with third parties. This could also lead to violations of user privacy [34].

Another risk is the lack of transparency in the algorithms employed by GAI. Not understanding how these algorithms function can lead to distrust in the decisions and recommendations provided by the devices and/or tools [35]. However, excessive reliance also poses risks. Overconfidence in these tools can hinder human creativity and innovation [36].

Another significant risk lies in the emotional and cognitive bonds that users may establish during the interaction process with GAI tools and/or devices. The integration of these technologies into our lives has involved them in personal and emotional aspects of people's lives. This raises the possibility that users may develop emotional dependence on machines, creating an affective connection that could influence traditional human relationships. This phenomenon emphasizes the importance of addressing the ethical and emotional aspects of Human-Generative Artificial Intelligence Interaction (HGAII onwards) [37].

Understanding how these technologies work, what they provide, how they can assist us, as well as the benefits they offer and the risks they pose, is essential for informed and responsible interaction. Promoting awareness and AI literacy enables users to make conscious decisions, critically evaluating the utility of these tools in their everyday lives. The key is to empower individuals with the necessary knowledge to harness the benefits of GAI ethically and equitably, while preserving autonomy and personal skills [15].

E. Expectations and Perceptions of GAI

The tools and devices of GAI have been introduced in society as a revolutionary technology with the potential to transform different areas of our lives. In fact, they have promised significant advancements in fields such as medicine, education, and art, among others [1]. The ability to generate written, visual, and auditory content has been perceived as a valuable contribution to saving time and resources, enabling professionals to focus on more complex tasks [15].

The discrepancy between expectations and reality in the interaction process with Conversational Agents (CAs onwards) highlights the complexity of the path to the full realization of GAI tools [38]. Clark et al. [25] found that we should not perceive this interaction process as an imitation of human capabilities but as a new process of communication and interaction. The expectation that GAI tools will come to comprehend their own existence and be capable of making independent decisions, surpassing predefined instructions, represents an ambitious horizon that has not yet been achieved [15]. A lack of knowledge about the current capabilities of these technologies can

impact user's risk perception [39], emphasizing the need for GAI literacy [15].

GAI-driven systems have limitations that prevent them from offering optimal responses indefinitely. The false hope that CAs are infallible negatively impacts the intention to use these tools when users encounter errors, especially in laboratory and field studies where the initial reliability rate was very high [40]-[42]. There is often an initial belief that these tools are error-free, generating high confidence. However, when users become aware of errors in responses or in understanding their requests, that confidence is weakened [33].

Trust has also been recognized as a factor predicting the quality of HRI and people's willingness to use social robots in certain tasks. The level of trust can be influenced by media representations, such as movies where robots dominate the world, creating often unrealistic expectations that may induce fear or rejection toward the adoption of these devices. However, what truly concerns users is the fear that social robots will replace human labor. In fact, positive attitudes have been found towards the presence of robots in jobs that require social skills [43].

III. Method

A. Methodology

The present research focuses on a pilot study. In this initial phase, a deliberately small sample has been selected for the questionnaire application, aiming to ensure the relevance and effectiveness of the items used. This pilot approach will allow for adjustments and refinements to the methodology before conducting large-scale data collection, ensuring the quality and validity of the results obtained.

B. Participants

A total of 100 young individuals aged between 18 and 34 participated. Specifically, 63.0% (*N*=63) of the participants are biologically male, while 37.0% (*N*=37) are biologically female. When inquiring about gender, 60.0% (*N*=60) identified with the male gender, while 38.0% (*N*=38) identified with the female gender. 2.0% (*N*=2) chose not to reveal this information. 99.0% (*N*=99) of the participants reside in Spain, with the majority of them residing in the Canary Islands (87.0%, *N*=87). 7.0% (*N*=7) reside in Madrid, 3.0% (*N*=3) in Andalucía, 1.0% (*N*=1) in Cantabria, and another 1.0% (*N*=1) in Galicia.

Regarding participant technology usage, 95.0% (*N*=95) of the participants use it to connect to the Internet, 79.0% (*N*=79) use it for communicative purposes, and 60.0% (*N*=60) use it for entertainment activities, such as gaming (shown in Table I).

TABLE I. Youth Internet Usage Habits

Items	Sex	1		2		3		4		5	
		N	%	*N*	%	*N*	%	*N*	%	*N*	%
I use technology to connect to the Internet	M	0	0	0	0	0	0	3	3	60	60
	F	0	0	0	0	0	0	3	3	34	34
I use technology to communi-cate	M	0	0	0	0	2	2	11	11	50	50
	F	0	0	0	0	0	0	8	8	29	29
I use technology to play	M	1	1	3	3	6	6	11	11	42	42
	F	1	1	3	3	4	4	11	11	18	18

a Items were assessed using a Likert scale, where 1 indicates completely disagree, and 5 indicates completely agree.

C. Research Questions

The aim of this pilot study is to identify both the expectations and the perceptions that young individuals have regarding GAI tools. Specifically, the following research questions will be addressed to achieve a comprehensive overview of the results:

- RQ1: To what extent are the youth familiar with the terminology of GAI?

- RQ2: What are the main concerns of the youth regarding GAI? Are there gender differences in these concerns?

- RQ3: How do youth perceive the effectiveness of GAI in communication and to what extent does it contribute to their well-being?

- RQ4: How do youth perceive GAI devices usefulness within the educational realm?

D. Data Collection

For data collection, an *ad hoc* questionnaire featuring two dimensions was employed. The first dimension, focused on participants' characterization, was designed to gather detailed information about participants. This dimension consists of 8 questions covering aspects such as sex, gender, age, place of residence, and technology-related habits. To explore participants' expectations and perceptions regarding GAI, a second dimension was designed, addressing key aspects such as concept understanding, frequency and types of use, as well as ethical considerations, privacy, and interaction. This dimension comprises a total of 46 items, of which 45 are assessed using a Likert scale ranging from 1 (completely disagree) to 5 (completely agree), and an additional optional qualitative item. The time required to complete the questionnaire ranged from 10-15 minutes, and it was filled out individually in digital format.

E. Procedure

The study employed a non-probabilistic sampling method, specifically targeting a population subgroup through the snowball sampling technique. To be more concrete, the questionnaires were distributed through email platforms and forums within the virtual campus of the University of La Laguna to students in the Bachelor's Degree in Computer Engineering, the Master's Degree in Teacher Training for Middle and High Education, Vocational Training and Language Training with a specialization in Technology Education, among other areas.

Once the deadline for completing the questionnaire expired, the data analysis process started. Initially, the results were recorded in a Google Sheet spreadsheet where participants' personal data were coded to ensure privacy. Subsequently, the template was imported into the IBM SPSS Statistics statistical software program to initiate the analysis and interpretation of the collected data.

F. Data Analysis

The analysis was conducted with a quantitative approach based on the research questions. Various analyses were employed in line with the defined objectives and variables.

In the initial phase, the internal consistency of the questionnaire was assessed through a factorial analysis, employing Cronbach's Alfa coefficient as a measure. The internal consistency of the questionnaire, as determined through the analysis, was found to be .93. Subsequently, descriptive analyses were conducted to explore participant's internet usage habits, employing cross-tabulations for this purpose (Refer to Table I). To examine variable distribution and adherence to normality, the Kolmogorov-Smirnov test was applied. Given the non-normal distribution of the majority of variables, non-parametric measures, concretely the Mann-Whitney U and independent samples t-tests,

were conducted. This comprehensive methodological approach facilitated a rigorous examination of various aspects of the research, offering a complete and detailed insight into the obtained results.

All analyses were conducted using the IBM Statistical Package for the Social Sciences (SPSS) software, version 29.0 for Windows.

IV. Results

A. Familiarity With GAI and Practical Application

Addressing the first research question, 20.0% (*N*=20) of the respondents indicated familiarity with the concept of GAI. However, when asked if they could define the concept, 24.0% (*N*=24) disagreed with being able, and 31.0% (*N*=31) neither agreed nor disagreed regarding understanding how GAI functions. Concerning practical usage, 23.0% (*N*=23) of the participants agree that they frequently use applications employing GAI. In the academic realm, 31.0% (*N*=31) stated that the use of GAI tools enhances their productivity in academic tasks. A 3.0% (*N*=3) mentioned that its use reduces productivity in these tasks, while it neither motivates nor demotivates 32.0% (*N*=32). In terms of future perspectives, 42.0% (*N*=42) firmly believe that GAI has significant potential to transform how we work and live. Regarding knowledge of usage, 33.0% (*N*=33) of participants assert that GAI tools record their request during the interaction, while 27.0% (*N*=27) express having no defined position on the matter. Concerning the learning capacity of these tools, 38.0% (*N*=38) agree that GAI tools learn and improve over time through interaction, and 34.0% (*N*=34) completely agree. In the context of ethical limits during interaction, 32.0% (*N*=32) remain neutral, neither agreeing nor disagreeing. On the other hand, 30.0% (*N*=30) agree, and 21.0% (*N*=21) completely agree that GAI tools have ethical limits during the interaction process. Additionally, 35.0% (*N*=35) agree that these tools have limited capacity to generate responses during interaction. Fig. 1 provides an overview of all results and items discussed.

B. Concerns and Considerations

Exploring the second specific research inquiry, 26.0% (*N*=26) of participants express a neutral stance, showing neither agreement nor disagreement regarding the respect for their privacy. On the other hand, 49.0% (*N*=49) voice significant concern, indicating full agreement that they are worried about the potential use of these tools to generate false or misleading content. In the realm of gender biases present in GAI tools, 18.0% (*N*=18) express concern, while another 24.0% (*N*=24) fully agree with this apprehension. In contrast, 24.0% (*N*=24) indicate being neither concerned nor unconcerned regarding this issue. Concerning transparency in the decision-making processes of GAI tools, 28.0% (*N*=28) adopt a neutral position, showing neither agreement nor disagreement with concerns about transparency. Conversely, another 28.0% (*N*=28) fully agree with concerns about the risk of developing a dependency on these tools in decision-making. Women demonstrate higher levels of concern across all evaluated aspects in comparison to men. However, the only significant gender difference in the analysis relates to their concerns about privacy when using GAI tools with a .024. Women seem to be more worried about keeping their personal information private while using these technologies compared to men. Fig. 2 displays the results for male respondents, while Fig. 3 presents the results for female respondents, facilitating a comparative analysis.

C. User Experience

Delving into the third targeted research questions, 39.0% (*N*=39) express neither agreement nor disagreement with feeling secure when interacting with GAI tools. In contrast, 7.0% (*N*=7) fully agree with this statement. Regarding the sense of companionship when interacting with these tools, 9.0% (*N*=9) agree, while 15.0% (*N*=15) neither agree nor
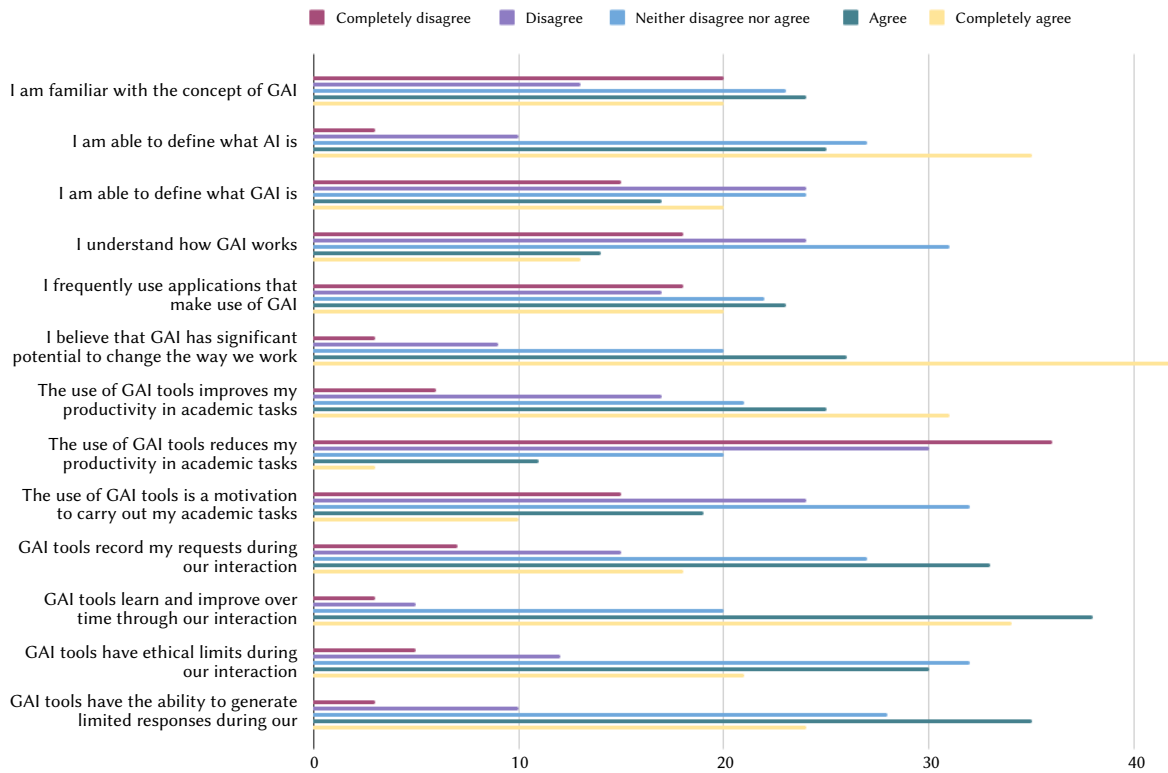
Fig. 1. Familiarity with GAI and practical application.
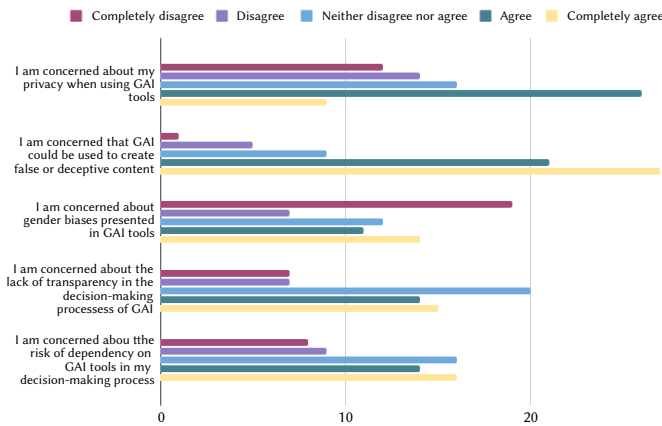
Fig. 2. Concerns and considerations of male participants.
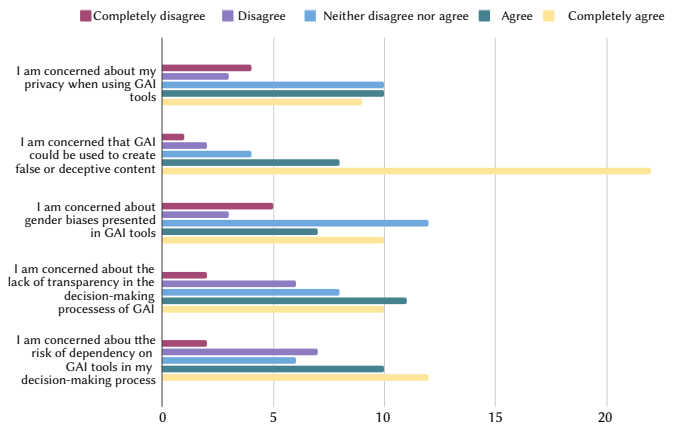
Fig. 3. Concerns and considerations of female participants.

disagree. Only 5.0% (*N*=5) agree that GAI understands their emotions, while a notable 49.0% (*N*=49) completely disagree with this statement. Concerning the responsiveness of GAI to emotions, only 6.0% (*N*=6) agree, while a significant 47.0% (*N*=47) completely disagree. Regarding overall satisfaction, 35.0% (*N*=35) agree with feeling satisfied with GAI tools. Additionally, 33.0% (*N*=33) agree that they can communicate effectively with GAI (Illustrated in Fig. 4). On the other hand, 26.0% (*N*=26) agree that they feel anxious when GAI tools fail to understand their requests. Furthermore, 25.0% (*N*=25) agree that GAI contributes to a positive experience. In emotional terms, 4.0% (*N*=4) completely agree, and 10.0% (*N*=10) agree that communication with GAI enhances their emotional well-being (Shown in Fig. 5).

D. Devices Employing GAI

Concerning the use of virtual assistants, 23.0% (*N*=23) express agreement when asked if the use of these devices increases their desire to learn new content. 37.0% (*N*=37) indicate no defined position

on whether virtual assistants foster a participative attitude towards learning, while 21.0% (*N*=21) agree. Regarding the impact on reflection on various topics, 18.0% (*N*=18) agree that virtual assistants encourage reflection. In terms of skills, and knowledge in the use of ICT, 31.0% (*N*=31) agree that these devices contribute to their development. In the realm of personalized learning, 31.0% (*N*=31) indicate having no clear stance on whether virtual assistants support this approach based on individual characteristics, while 25.0% (*N*=25) agree. Finally, 18.0% (*N*=18) concur that virtual assistants facilitate control and evaluation of the learning process.

In the realm of chatbot usage, 24.0% (*N*=24) of participants agree that using chatbots increases their motivation to learn new content. In terms of promoting a participative attitude towards learning, 18.0% (*N*=18) agree with the contribution of chatbots. Concerning the ability to foster reflection on various topics, 29.0% (*N*=29) agree, while another 29.0% (*N*=29) adopt a neutral position, neither expressing
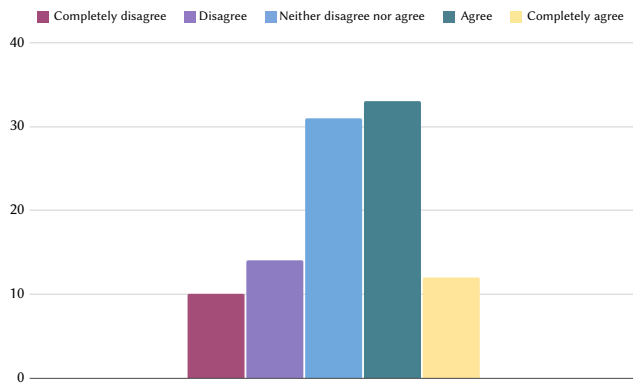
Fig. 4. Results of effectively communicating with GAI tools.

Fig. 5. Outcomes of dialogue with GAI to enhance emotional wellbeing.

agreement nor disagreement. In the aspect of developing skills and knowledge in the use of ICT, 40.0% (*N*=40) confirm that chatbots play a positive role in this process. In terms of personalized learning, 26.0% (*N*=26) agree that chatbots support this approach based on individual characteristics. Furthermore, an equal percentage concur that chatbots facilitate control and evaluation of the learning process (26.0%, *N*=26).

Regarding the use of social robots, 30.0% (*N*=30) express total disagreement with the statement that using these devices increases their motivation to learn new content. On the other hand, 19.0% (*N*=19) agree that social robots foster a participative attitude during the learning process. In terms of promoting reflection on various topics, 18.0% (*N*=18) agree, while 24.0% (*N*=24) disagree with this statement. In the aspect of developing skills and knowledge in the use of ICT, 35.0% (*N*=35) neither express agreement nor disagreement, while 22.0% (*N*=22) agree, and 12.0% (*N*=12) fully agree. Concerning support for personalized learning based on individual characteristics, 43.0% (*N*=43) neither express agreement nor disagreement, and 16.0% (*N*=16) agree. In terms of facilitating control and evaluation of learning, 35.0% (*N*=35) neither show agreement nor disagreement, while 15.0% (*N*=15) agree.

In response to the fourth research question, it is observed that young individuals find virtual assistants and chatbots more useful than social robots in the educational context.

V. Discussions

The aim of this pilot study was to identify the expectations and perceptions that young students have regarding GAI tools. The following four research questions were formulated to facilitate a comprehensive analysis and understanding within this specific domain: RQ1. To what extent are the youth familiar with the terminology of GAI?, RQ2. What are the main concerns of the youth regarding GAI? Are there gender differences in these concerns?, RQ3. How do youth perceive the effectiveness of GAI in communication and to what extent does it contribute to their well-being?, and RQ4. How do youth perceive GAI devices usefulness within the educational realm?

The intersection between familiarity and practical usage of GAI reveals insights into the adoption of emerging technologies, reaffirming the necessity of technical understanding for effective implementation [15]. Despite a limited understanding of GAI terminology and mechanics, a significant portion of respondents actively engage with GAI tools, highlighting a disconnect between knowledge and usage. This paradox suggests that user-friendly interfaces outweigh the need for in-depth technical comprehension, fostering widespread acceptance among university students. However, while the accessibility of these interfaces facilitates the use of GAI tools without requiring technical expertise, it simultaneously omits the potential risks and vulnerabilities that students may encounter
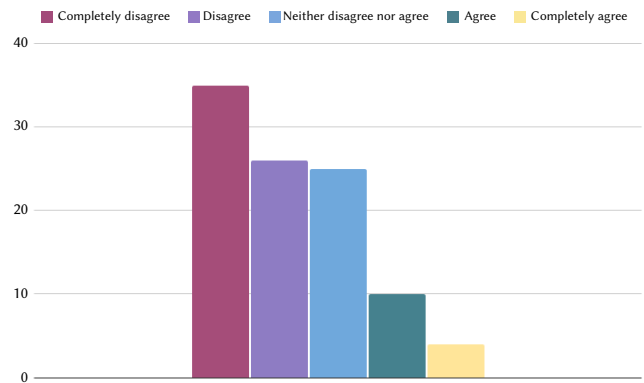
during the interaction processes with GAI tools.

Fig. 1 illustrates a strong belief among respondents that GAI tools can transform the way we work and enhance productivity in academic tasks. This belief is evident in the high percentage of participants who agree and completely agree that GAI can significantly influence and improve different aspects of professional and educational activities. While the belief in the transformative potential of GAI tools appears promising for their continued implementation in both professional and academic context, there are concerns regarding the respondents' confidence in these tools without adequate consideration of potential gender biases, lack of transparency in decision-making processes, and the risk of dependency on these tools for decision-making. This observation prompts us to question the future of GAI tools if upcoming computer scientists fail to address the fundamental aspects of these technologies. Ignoring critical issues such as potential biases and transparency in decision-making could weaken trust and minimize their effectiveness over time. Additionally, the lack of concerns among respondents regarding their dependence on GAI tools raises questions about autonomy and judgment in an increasingly AI-driven world. While the convenience of these tools is clearly attractive, it is crucial to approach the risk of excessive reliance with caution, as it could adversely affect critical thinking and informed decision-making [14]. Women exhibited higher levels of concern across all items related to the potential risks of GAI tools. As illustrated in Fig. 3, this heightened concern is evident in their responses. However, there is only one item with a statistically significant difference between genders: "I am concerned about my privacy when using GAI tools". Women demonstrate a markedly greater concern about this issue than men (Shown in Fig. 2). Conversely, the level of consensus among respondents expressing concern about using GAI tools for creating false or deceptive content is striking. This highlighted item reflects a shared sensitivity among participants, suggesting a widespread perception of the potential threat that GAI tools pose to information integrity. This collective concern may be attributed to prior experiences with online misinformation, manipulation campaigns, and increased awareness of the potential malicious uses of technology or the fear of plagiarism.

In the context of user experience, the natural interaction offered by GAI-based tools has created a new communication process with technology. The ability to understand natural language, interpret voice commands, and adjust to individual preferences has expanded comfort and accessibility for users [25]. The results of user communication are presented in Fig. 4. However, the findings related to emotional well-being, highlighted in Fig. 5, are less encouraging. Few users feel accompanied when interacting with GAI tools, and even fewer believe that GAI understands or responds appropriately to their emotions. This could be attributed to a gap in the ability of GAI tools to establish affective emotional connections or issues in the communication process itself [15]. There is a moderate level of agreement among users in terms

of effective communication. It is important to consider that users' preconceived perceptions of a machine's inability to enhance their well-being may influence these low values. Similar to the emotional connections established in childhood with toys, we may also develop emotional bond with GAI devices, especially when these devices provide personalized responses and facilitate various tasks [37].

Investigating the variety of GAI devices currently used by young individuals provides valuable insights, particularly regarding their utility in education. The results show that participants perceive virtual assistants and chatbots as more beneficial than social robots in the educational context. Virtual assistants are widely perceived to support personalized learning effectively and chatbots are similarly valued. However, social robots are seen as less contributive to personalized learning. This disparity could be due to the greater accessibility and ease of use of virtual assistants and chatbots compared to the more complex implementation of social robots in educational settings. The preference for virtual assistants and chatbots likely stems from their ability to provide quick, personalized responses [23]-[25], and their adaptability to various platforms and devices. These findings highlight that for young individuals, immediate utility and efficiency are critical in evaluating the tools applications of GAI in education.

VI. Conclusions

This study sheds light on the expectations and perceptions of young individuals regarding GAI. Our findings reveal a nuanced landscape where familiarity with GAI terminology and detailed understanding of its internal mechanisms are limited, yet there is a notable adoption of these tools in daily practice. The accessibility and practical benefits of GAI applications appear to drive their widespread acceptance despite the technical knowledge gap. A significant portion of respondents believed that GAI tools have the potential to transform the way we work, with many recognizing the enhancement of productivity in academic tasks. This optimistic outlook underscores the expectations placed on GAI technologies. However, there is a notable lack of concern about privacy, transparency in decision-making, and dependency on these tools. Women expressed higher levels of concern across all items related to potential risks, with privacy being a significant gender-specific issue. Moreover, the study highlights a critical awareness among respondents about the risk of GAI tools generating false or deceptive content, reflecting a shared sensitivity to information integrity. In the realm of user experience, GAI tools are appreciated for their natural interaction capabilities, which enhance comfort and accessibility. However, the emotional connection with these tools remains weak, indicating a gap in their ability to establish affective bonds with users. When it comes to educational applications, virtual assistants and chatbots are perceived as more useful than social robots. In conclusion, young individuals demonstrate a multifaceted relationship with GAI, marked by high expectations for its transformative potential and practical benefits, alongside a notable lack of concern regarding privacy, transparency, and misinformation. These insights indicate that, while considerable enthusiasm exists for integrating GAI into daily life, addressing these critical concerns will be essential for ensuring these technologies' sustainable and ethical employment.

The results presented in this study should be interpreted with caution as they are not generalizable due to their adaptation to a specific context and a small sample size. It is important to note that this study has been designed as a pilot, implying the need to assess which items are truly significant for the research and which are not, thus adjusting the focus for future investigations. Additionally, the study has certain limitations that need to be considered. Some participants expressed difficulties in answering certain questions due to a lack

of knowledge about the subject, especially regarding the use and understanding of devices such as social robots. These challenges were directly communicated to the researchers. Furthermore, the uneven participation of men and women may influence the conclusions, given the study's majority representation of men, possibly attributable to the choice of fields with low female presence, such as computer engineering, and the technological specialization of the teacher training master's program. In future research, it is recommended to include the field of study in the questionnaire to identify the background of female participants and determine if they all come from technological fields or not. These limitations should be considered when interpreting the findings and provide opportunities for improvement and refinement of methodology in subsequent research.

The current research sets the stage for future lines of exploration. To delve deeper into user perspectives on GAI tools, exploring the underlying reasons behind users' deeply rooted beliefs about their usage is essential. Investigating the sources from which users access information and form opinions regarding GAI tools could offer valuable insights into the shaping of their attitudes. Furthermore, exploring the long-term impact and evolving role of GAI in educational settings could be another promising trajectory, shedding light on its efficacy over time and potential implications for pedagogical approaches. Additionally, investigating strategies to address the identified concerns, such as privacy, emotional engagement, and gender disparities, could contribute to the development of more inclusive and effective GAI tools. Comparative studies across different demographic groups and cultural contexts could also offer valuable insights into the context-specific nature of the observed trends, specifically when comparing the usage of GAI devices (e.g., virtual assistants, chatbots, and social robots). By addressing these aspects, future research can contribute to a comprehensive understanding of the multifaceted dynamics surrounding the integration of GAI in higher education and its broader implications. The study also highlights the importance of understanding the culture of AI, which plays a significant role in shaping how individuals perceive and interact with GAI tools. Future research should investigate how these cultural attitudes influence users' expectations and concerns about GAI, and how these perceptions vary across different demographics and cultural contexts.

References

[1] K.B. Ooi et al., "The potential of Generative Artificial Intelligence across disciplines: Perspectives and future directions", *Journal of Computer Information Systems*, no. 1, pp. 1-32, 2023, doi: 10.1080/08874417.2023.2261010

[2] M. Mandapuram, S. Surjan-Gutlapalli and M. Reddy, "Investigating the prospects of Generative Artificial Intelligence", *Asian Journal of Humanity, Art and Literature*, vol. 5., no. 2, pp. 167-174, 2018, doi:

10.18034/ajhal.v5i2.659

[3] S. Peña-Fernández, U. Peña-Alonso and M. Eizmendi-Iraola, "El discurso de los periodistas sobre el impacto de la Inteligencia Artificial Generativa en la desinformación", *Estudios sobre el Mensaje Periodístico*, vol. 29, no. 4, pp. 833-841, 2023, doi: 10.5209/esmp.88673

[4] N. Rane, "Roles and Challenges of ChatGPT and similar Generative Artificial Intelligence for achieving the Sustainable Development Goals (SDGs)", *SRRN*, pp. 1-14, 2023, doi: 10.2139/ssrn.4603244

[5] P. Zhang and M.N. Kamel-Boulos, "Generative AI in medicine and healthcare: Promises, opportunities and challenges", *Future Internet*, vol. 15, no. 9, pp. 1-15, 2023, doi: 10.3390/fi15090286

[6] J. Qadir, "Engineering education in the era of ChatGPT: Promise and pitfalls of Generative AI for education", *IEEE Global Engineering Education Conference*, Kuwait, Kuwait, 2023, pp. 1-9, doi: 10.1109/EDUCON54358.2023.10125121

[7] M.A. Ali-Elfa and M.E. Tawfilis-Dawood, "Using Artificial Intelligence for enhancing human creativity", *Journal of Art, Design and Music*, vol. 2, no. 2, pp. 106-120, 2023, doi: 10.55554/2785-9649.1017

[8] M.V. Roehling, "An empirical assessment of alternative conceptualizations of the psychological contract construct: Meaningful differences or "much to do about nothing?", *Employee Responsibilities and Rights Journal*, vol. 20, no. 4, pp. 261-290, 2008, doi: 10.1007/s10672-008-9085-z

[9] F. Ederer, "Trust and promises over time", *American Economic Journal: Microeconomics*, vol. 14, no. 3, pp. 304-320, 2022, doi: 10.1257/mic.20200049

[10] P. Budhwar et al., "Human resource management in the age of Generative Artificial Intelligence: Perspectives and research directions on ChatGPT", *Human resources management Journal*, vol. 33, no. 3, pp. 606-659, 2023, doi: 10.1111/1748-8583.12524

[11] M. Borup, N. Brown, K. Konrad and H. Van-Lette, "The sociology of expectations in science and technology", *Technology Analysis & Strategic Management*, vol. 18, no. 3, pp. 285-298, 2006, doi: 10.1080/09537320600777002

[12] COMEST, "Preliminary study on the ethics of Artificial Intelligence", UNESCO, Paris, 2019.

[13] M. Haenlein and A. Kaplan, "A brief history of AI: On the past, present and future of Artificial Intelligence", *California Management Review*, vol. 61, no. 4, pp. 5-14, 2019, doi: 10.117/0008125619864925

[14] UNESCO, "K-12 AI curricula: a mapping of government endorsed AI curricula", UNESCO, France, 2022.

[15] A. Pedreño-Muñoz, R. González-Gosálbez, T. Mora-Illán, E.M. Pérez-Fernández, J. Ruiz-Sierra and A. Torres-Penalva, "La Inteligencia Artificial en las universidades: Retos y oportunidades. Informe anual sobre IA y educación superior", 2024.

[16] F.J. García-Peñalvo, F. Llorens-Largo, and J. Vidal, "The new reality of Education in the face of advances in Generative Artificial Intelligence", *Revista Iberoamericana de Educación a Distancia*, vol. 27, no. 1, pp. 9-39, 2023, doi: 10.5944/ried.27.1.37716

[17] S. Murugesan and A. Kumar-Cherukuri, "The rise of Generative Artificial Intelligence and its impact on education: The promises and perils", *Computer*, vol. 56, no. 5, pp. 116-121, 2023, doi: 10.1109/MC.2023.3253291

[18] Y. Cao et al., "A comprehensive survey of AI-Generated Content (AIGC): A history of Generative AI from GAN to ChatGPT", *Journal of the ACM*, vol. 37, no. 4, pp. 1-44, 2023, doi: 10.48550/arXiv.2303.04226

[19] I. Vaccari, V. Orani, A. Paglialonga, E. Cambiaso and M. Mongelli, "A Generative Adversarial Network (GAN) technique for internet of medical things data", *Sensors*, vol. 21, no. 11, pp. 1-14, 2021, doi: 10.3390/s21113726

[20] N. Rane, "Role and challenges of ChatGPT and similar Generative Artificial Intelligence in business management", *SSRN*, pp. 1-12, 2023, doi: 10.2139/ssrrn.4603227

[21] D. Baidoo-Anu and L. Owusu-Ansah, "Education in the era of Generative Artificial Intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning", *Journal of AI*, vol. 7, no. 1, pp. 52-62, 2023, doi: 10.61969/jai.1337500

[22] N. Anantrasirichai and D. Bull, "Artificial Intelligence in the creative industries: A review", *Artificial Intelligence Review*, vol. 55, pp. 589-656, 2022, doi: 10.1007/s10462-021-10039-7

[23] T. Sakirin and R. Ben-Said, "User preferences for ChatGPT-powered conversational interfaces versus traditional methods", *Mesopotamian Journal of Computer Science*, vol. (2023), pp. 24-31, 2023, doi: 10.58496/

MJCSC/2023/004

[24] S. Malodia, N. Islam, P. Kaur and A. Dhir, "Why do people use Artificial Intelligence (AI)-Enabled Voice Assistants?", *IEEE Transactions on Engineering Management*, vol. 17, pp. 491-505, 2021, doi: 10.1109/TEM.2021.3117884

[25] Clark et al., "What makes a good conversation? Challenges in designing truly Conversational Agents", *in CHI Conference on Human Factors in Computing Systems*, Glasglow, UK, 2019, pp. 1-12, doi: 10.1145/3290605.3300705

[26] A.P. Chaves and M.A. Gerosa, "How should my chatbot interact? A survey on social characteristics in Human-Chatbot Interaction design", *International Journal of Human-Computer Interaction*, vol. 37, no. 8, pp. 729-758, 2020, doi: 10.1080/10447318.2020.1841438

[27] C.S. González-González, V. Muñoz-Cruz, P.A. Toledo-Delgado and E. Nacimiento-García, "Personalized gamification for learning: A reactive chatbot arquitecture", *Sensors*, vol. 23, n°. 1, pp. 1-18, 2022, doi: 10.3390/s23010545

[28] A. Henschel, G. Laban and E.S. Cross, "What makes a Robot social? A review of Social Robots from science fiction to home or hospital near you", *Current Robotics Reports*, vol. 2, pp. 9-19, 2021, doi: 10.1007/s43154-020-00035-0

[29] C.S. González-González, R.M. Gil-Iranzo and P. Paderewski-Rodríguez, "Human-Robot Interaction and sexbots: A systematic review", *Sensors*, vol. 21, n°. 1, pp. 2-18, 2021, doi: 10.3390/s21010216

[30] Z. Bahroun, C. Anane, V. Ahmed and A. Zacca, "Transforming education: A comprehensive review of Generative Artificial Intelligence in educational settings through bibliometric and content analysis", *Sustainability*, vol. 15, no. 17, pp. 1-40, 2023, doi: 10.3390/su151712983

[31] M.U. Hadi et al., "Large language models: A comprehensive survey of its applications, challenges, limitations and future prospects", *ResearchGate*, Preprint.

[32] H. Ibrahim et al., "Perception, performance, and detectability of conversational artificial intelligence across 32 university courses", *Scientific Reports*, vol. 13, n°. 12187, pp. 1-13, 2023, doi: 10.1038/s41598-023-38964-3

[33] I. Amaro, A. Della-Greca, R. Francese, G. Tortora and C. Tucci, "AI unreliable answers: A case study on ChatGPT", *in International Conference on Human-Computer Interaction*, Copenhagen, Denmark, 2023, pp. 23-40, doi: 10.1007/978-3-031-35894-4

[34] P. Won-Kim, "A framework to overcome the dark side of Generative Artificial Intelligence (GAI) like ChatGPT in social media and education", *IEEE Transactions on Computational Social Systems*, pp. 1-9, 2023, doi: 10.1109/TCSS.2023.3315237

[35] S. Larsson and F. Heintz, "Transparency in Artificial Intelligence", *Internet Policy Review*, vol. 9, no. 2, pp. 1-16, 2020, doi: 10.14763/2020.2.1469

[36] N. Rane, "ChatGPT and similar Generative Artificial Intelligence (AI) for smart industry: Role, challenges and opportunities for Industry 4.0, Industry 5.0 and Society 5.0", *SRRN*, pp. 1-12, 2023, doi: 10.2139/ssrn.4603234

[37] A.E. Cotino-Arbelo, C.S. González González, and J.M. Molina-Gil, "Embracing the future: Unveiling the revolution of Human-AI Interaction in the digital education era", *in XIII International Conference on Virtual Campus*, Porto, Portugal, 2023, to be published.

[38] E. Luger and A. Sellen, "Like having a really bad PA: The gulf between user expectation and experience of Conversational Agents", *in CHI Conference on Human Factors In Computing Systems*, New York, NY, USA, 2016, doi: 10.1145/2858036.2858288

[39] T. Brown et al., "Language models are few-shot learners", *in Conference on Neural Information Processing Systems*, Vancouver, Canada, 2020, pp. 1877-1901, doi: 10.48550/arXiv.2005.14165

[40] S.F. Rohden and D. García-Zeferino, "Recommendation agents: An analysis of consumers' risk perceptions toward artificial intelligente", *Electronic Commerce Research*, vol. 23, pp. 2035-2050, 2023, doi: 10.1007/s10660-022-09626-9

[41] A. Glass, D.L. McGuiness and M. Wolverton, "Toward establishing trust in adaptive agents", *in 13th International Conference on Intelligent User Interfaces*, Gran Canaria, Spain, 2008, pp. 227-236, doi: 10.1145/1378773.1378804

[42] S. Moran et al., "Team reactions to voiced agent instructions in a pervasive game", *in International Conference on Intelligent User Interfaces*,

Santa Monica, CA, USA, 2023, pp. 371-382, doi: 10.1145/2449396.2449445

[43] S. Naneva, M. Sarda-Gou, T.L. Webb and T.J. Prescott, "A systematic review of attitudes, anxiety, acceptance, and trust towards Social Robots", *International Journal of Social Robotics*, vol. 12, pp. 1179-1201, 2020, doi: 10.1007/s12369-020-00659-4

Andrea E. Cotino Arbelo

Currently pursuing a Ph.D. in industrial, computer and environmental engineering at the University of La Laguna, Spain, Andrea holds a degree in primary education from the same university in 2021 and completed her master's in research and innovation in curriculum and training at the University of Granada, Spain, in 2022. Navigating the academic landscape with a keen focus, she explores the transformative potential of generative artificial intelligence, seeks to enhance user experience, and advances educational digital methodologies. A member of the Interaction, ICT and Education (ITED) research group and the University Institute of Women's Studies since 2023, Andrea actively participates in multiple projects, conferences, and publications within the field.

Carina S. González González

Is a full professor of computer architecture and technology at the University of La Laguna, Spain. She obtained her Ph.D. in computer science from the University of La Laguna in 2001, and a Ph.D. in Social Sciences and Education from the University of Huelva, Spain, in 2020. Her research focuses on the application of AI techniques and accessible, intelligent interfaces in education within the Department of Computer Engineering and Systems. She serves as the Director of the Institute of Women's Studies and leads the research group Interaction, ICT, and Education (ITED). Additionally, she directs the Interactive Digital Culture Classroom and has held the position of Director of Innovation and Educational Technology at the University of La Laguna during various periods (2011; 2015-2019).

Jezabel Molina Gil

Is an assistant professor of computer science and artificial intelligence at the University of La Laguna, Spain. She received her computer science engineering degree from the University of Las Palmas de Gran Canaria in 2007 and her PhD from the University of La Laguna in 2011. Her research is focused on VANET security, specifically in cooperation and data aggregation. She belongs to the CryptULL research group devoted to the development of projects on cryptology (since 2007), and is involved in several projects and publications related to this area. She has authored several conference and journal papers.

Gaming as a Medium for the Expression of Citizens' Views on Environmental Dilemmas

Dai Griffiths[1]*, Jude Ower[2], Paul Hollins[3], Anchal Garg[3]

[1] Research Institute for Innovation & Technology in Education (UNIR iTED), Universidad Internacional de La Rioja, Logroño, La Rioja (Spain)
[2] PlanetPlay (UK)
[3] The University of Bolton (UK)

* Corresponding author: david.griffiths@unir.net

Abstract

The decline of traditional media and channels of communication has led to policymakers experiencing difficulty in understanding public sentiment. A case study was conducted to explore how games-based activities can be used to provide a link between citizens and policy makers. A system developed by PlanetPlay, and extended in the GREAT project, was used to embed a survey in the game SMITE. The intervention and survey questions were designed in collaboration with the United Nations Development Programme (UNDP) and the Hi-Rez game studio. The effectiveness of the infrastructure and the collaborative approach were demonstrated. The results revealed some significant differences in views on climate change between different age groups, genders, and education level. However, the data was heavily skewed towards males in the 18-35 age group, and to respondents in the United States, which limited the generalizability of the findings. It was concluded that in-game placement in collaboration with games studios is more effective than paid placement, and that a wider variety of games is needed to ensure that a study has an adequate range of respondent profiles. Finally, reflections are offered on the possible role of artificial intelligence in gathering such data.

I. Introduction

THE research reported here was carried out in the context of two interconnected cultural trends. The first, the increasing creative range and reach of the games industry has been generally welcomed. In contrast the second trend, the decline of traditional media and channels of communication has given widespread cause for concern, particularly as regards channels of communication linking citizens with civic authorities and policymakers. There may or may not be a degree of causal relationship between the two processes, but that is not our concern here. Rather, we identify and explore an opportunity to make use of the former in addressing some of the concerns raised by the latter. Before describing our study, we briefly introduce these two trends.

A. The Decline of Traditional Communication Channels

There is extensive evidence to support the statement by Contreras-Espinosa and Blanco [1] that "many democracies are facing, as a growing problem, a breach of communication between citizens and their political representatives". Since the 1990s, the proportion of citizens who are "dissatisfied" with democracy in their countries has risen by almost 10 percentage points globally, and the deterioration has been particularly marked in high-income, "consolidated" democracies, where the proportion has risen to a third to half of all citizens [2]. Similarly, the United Nations [3] considers that distrust of news sources and scientists is at an all-time low. Dissatisfaction and mistrust correlate with skepticism, for example concerning vaccines and covid19 [4] [5]. As Morelli has argued, these lower levels of trust in markets, governments, and political institutions have led to a crisis among traditional parties, and to an associated rise of populist rightwing parties [6].

The causes of this change are complex and contested, but it is relevant that there has been a marked decadence of institutions which have traditionally served to channel citizens' views to policymakers. The International Labor Organization has reported that the past thirty or forty years have been marked by the replacement of older unionized workers with less unionized but better educated younger workers [7]. Similarly, and despite high-profile fundamentalist exceptions, "most high-income countries show *declining* emphasis on religion" [8] (p.79, emphasis in the original). Perhaps most dramatic is the world-wide transformation of the news media landscape. For example, in the United States newspaper circulation has fallen by about two thirds since 1990

[9] and, correspondingly, 86% of US adults often or sometimes get news from a smartphone, computer or tablet, with 56% who doing so often [10]. This transformation no doubt reflects the greater convenience of digital distribution but is also impacted by two additional factors. Firstly, some politicians have undermined news media which they see as opposed to their interests, most notably Donald Trump, whose "antagonistic tweets are a systematic approach to delegitimize the news as an institution" [11]. Secondly, as Kreps et al. [12] observe, generative AI is flooding the media with enormous volumes of content, which is usually of little value and may constitute misinformation. They argue that "this undermines efforts to understand constituent sentiment, threatening the quality of democratic representation". It is precisely these difficulties in understanding public sentiment, flowing from the range of factors that we have mentioned, which motivate the research reported here.

B. The Emergence of Game Culture

In parallel with the decline of traditional news media, there has been a radical transformation in entertainment media, with the digital games industry having become a larger economic sector than either music or film, and which has been estimated to have generated $227 billion worldwide in 2023 [13]. It can hardly be doubted that this major cultural change has had a major impact on society, but there is no consensus on the consequences, nor even if these are positive or negative. According to Adrienne Shaw [14], the term 'video game culture' became common during the first decade of the 21st Century. However, she also warns that games players are highly varied in age, genders, sexualities, races, religions and nationalities, and that "Not all of these types of play and players can be encompassed in a study of an isolated gamer community" [14].

What, then constitutes the game (or games, or gaming) culture that has been a focus of so much discussion over the past fifteen years? Participants in games culture cannot readily be identified, as they do not usually wear distinctive clothing, despite the popularity of 'cosplay' on special occasions, as described by Lunning [15]. Mia Consalvo argues that membership of games culture is not only, or even mainly, concerned with playing games, it is marked by being knowledgeable about games, passing that information on, and having opinions about games [16].

Games culture has often been critiqued for being male dominated, aggressive and sexist, and, for example, Vergel et al. conclude that cybersexism and its manifestations "are a harsh reality for women who want to play digital and online games" [17]. Be this as it may, it is noteworthy that many gamers are women (though it is not clear if this undermines the conclusion of Vergel et al., or if it makes it still more concerning). In Europe 2022 46.7% of video game and console players were women, while for smartphone and tablet games they were in the majority (51%) [18]. There has also been widespread concern that aggressive video games are fomenting violence in society, and especially among people, but there is continuing doubt about the reality of this impact. For example, in 2020 Drummond et al. reported that "meta-analytic studies now routinely find that the long-term impacts of violent games on youth aggression are near zero" [19], while in a meta-analysis of 2021 Burkhardt and Lenhard identified "a significant and meaningful positive effect of VVG on subsequent physically aggressive behavior" [20].

From a more practical perspective, gaming, together with social media, web browsing, occupies a large amount of young people's time, with one study estimating U.S. teens' screen time at 8.39 hours per day, excluding educational activities [21]. There is concern that this highly competitive attention economy [22] may result in 'attentional serfdom' [23], and a paralysis of political participation in the face of the dilemmas which face society.

II. Motivation, Objectives and Research Questions

A. The Motivation for This Study

As Kroger et al. argue [24], the prevailing business model of the games industry is increasingly dependent on harvesting and making use of personal data for competitive advantage, often without players being aware of what data is being collected or for what purposes. The rapid introduction of artificial intelligence (AI) tools serves to accelerate this trend and make it still more opaque. Without entering into the rights and wrongs of this practice, we seek to show that another approach to data gathering through games is possible, with players choosing to provide data relating to issues which are relevant to them. Moreover, we position this data-gathering in terms of open science, proposing a methodology within which the participation of stakeholders can be maximized, and data can be made widely available for social benefit, in our case the expression of citizens' views on policy dilemmas.

The strength of games culture has long been seen as an opportunity to communicate ideas, promote attitudinal change, and enhance educational processes. However, there is a mismatch between the promise of these approaches and the disappointing scale of practical achievements. Moreover, while the Council of Europe has identified "great potential of video games in promoting positive cultural and social changes" [25], social and educational applications of games usually involve reception by the player of ideas or knowledge, and do not engage players in building an inquiry or making a contribution to society. Within this context, the work reported here addresses a gap in the research literature: methods are not described whereby participation in gaming culture can enable citizens to express their attitudes and preferences, and so address the problem of understanding constituent sentiment, identified by Kreps et al., above.

The case study described here leans heavily on the infrastructure and business processes of PlanetPlay (originally developed by Playmob, who were acquired by PlanetPlay in 2024). An additional motivation for this study is to examine the potential of these tools in the context of an academic case study, for the first time.

B. Objectives and Research Questions

In line with this motivation, our overarching research objective was to understand how games-based activities can be used to provide a link between citizens and policy makers. To this end, the practical objective of the work carried out in the case study was to provide insight to the participating policy stakeholders about citizens' views on a range of climate goals. In addressing our research objective, we sought to answer three of the wider research questions which have been defined for the GREAT project within which this research was embedded:

RQ1 Which games-based activities can be used to elicit, represent and communicate citizens' views on policy dilemmas?

RQ2 How effective are games-based activities in eliciting, representing and communicating citizens' views on policy dilemmas?

RQ3 How efficient is the use of games-based activities in eliciting, representing and communicating citizens' views on policy dilemmas?

III. Related Work

Games have been used for many years as an educational resource, and a substantial body of research has investigated its impact on different fields, as summarized, for example the systematic reviews offered by Yu et al. [26] for online education, Guan et al. [27] for primary education, and Vlachopoulos and Makri [28] for schools. More

specifically, games have long been used to enhance awareness and understanding of environmental and other social issues. For example a systematic review by Janakiraman et al., [29] argues that games have demonstrated the potential for producing attitudinal change, while Dhiman [30] concludes that games can "educate, advocate, create empathy, and build communities around social issues". This large body of research forms the background to our work but does not directly inform our research objective. However, the use of 'gamification' to gather citizens' views is more immediately relevant.

Karl M. Kapp merged a number of definitions of gamification in describing it as "...using game-based mechanics, aesthetics and game thinking to engage people, motivate action, promote learning, and solve problems" [31] (p.10). However, this can refer to a wide variety of approaches. In their systematic review, Keusch and Zhang [32] observe that forms of gamification in surveys vary widely, including "simply rephrasing questions to sound more game-like ... virtual badges and other intangible social rewards, and embedding the entire survey experience into a game where respondents are assigned to avatars and adventure through a fantasy land as they answer survey questions". Following Hamari, Koivisto, and Sarsa [33] they refer to such game elements as 'motivational affordances'.

Building on a widely adopted classification proposed by Yee [34], Blanco et al. reviewed the use of such elements in e-government services, and distinguish three relevant categories of gamification mechanics and game-design features: immersion (e.g. storytelling, avatars or role-play); achievement-related (e.g. challenges, badges, leaderboards or progression metrics); and social (e.g. social interaction and collaboration) [35].

In addition to the different categories of gamification that can be applied, it is important to consider the types of citizen engagement which they can support. Arnstein [36] made an early contribution to this discussion, conceiving of a ladder of participation consisting of three stages: non-participation, tokenism and genuine participation, each of which has multiple rungs. Mayer [37] also identifies three levels of relationship between citizens and political entities.

> E-enabling is about supporting those who would not typically access the internet and take advantage of the large amount of information available. ...
>
> E-engaging with citizens is concerned with consulting a wider audience to enable deeper contributions and support deliberative debate on policy issues. The use of the term 'to engage' in this context refers to the top-down consultation of citizens by government or parliament. ...
>
> E-empowering citizens is concerned with supporting active participation and facilitating bottom-up ideas to influence the political agenda. ... Here there is recognition that there is a need to allow citizens to influence and participate in policy formulation.

In a similar vein, Thiel et al. [38] distinguish between one-way and two-way communication in gamified participation approaches, corresponding roughly to the second and third of Mayer's three categories.

Three challenges for gamified surveys can be distinguished in the literature. Firstly, the expected improvement in engagement levels has not materialized, and Gastil and Broghammer write that "Unrealistic expectations are common when government and civic organizations adopt digital technologies to improve public engagement" [39]. This is true for the gamification of surveys, where evidence for a transformative impact is scant. A systematic review by Oliveira and Paula on this topic concludes that "it is not possible to say whether gamification stimulates engagement", but adds that there are indications that gamified surveys are more attractive and easier to answer [40].

Secondly, methodological concerns have been raised. In their review of gamified surveys Keusch and Zhang discuss the risk of "potential

bias as a result of making surveys fun":

> The biggest issue about survey gamification still concerns the influence of gamification on measurement error. One major challenge is that gamifying surveys often involves using techniques, such as rewording a question, changing response format (e.g., drag and drop), and adding additional visual elements, all of which inherently affect traditional data quality measures, such as response times, straightlining, and length of open-ended questions. [32]

Thirdly, as Harms et al. commented "survey gamification requires a lot of effort" [41], and it is not clear that the benefits are commensurate with this effort.

We return to these three challenges when we discuss our conclusions.

IV. Method and Tools

A. Study Design

The GREAT project has developed a case study methodology for use in a series of case studies linking citizens and policy stakeholders through games-based activities. This is designed as a cycle of steps, shown in Fig. 1. The expected activities and outputs of each step in the cycle are described in detail in GREAT deliverable D4.2 [42], together with templates for planning and documentation and guidance for the leaders of case studies. An evaluation framework has been developed, with a set of instruments for use in the design of individual case studies [43]. The steps are not mandatory, given the range of requirements created by the use of two contrasting platforms (PlanetPlay surveys and in-depth exploration of dilemmas in serious games using SGI's DiBL platform[1]), to address a wide range of institutional and geographical contexts. In the present paper we report on steps one to six, which correspond to our focus on the design, implementation and analysis of the embedded survey. The study is exploratory, in the sense that it investigates the potential of the approach and seeks to clarify the most effective use that can be made of the infrastructure in planned GREAT studies at a larger scale.
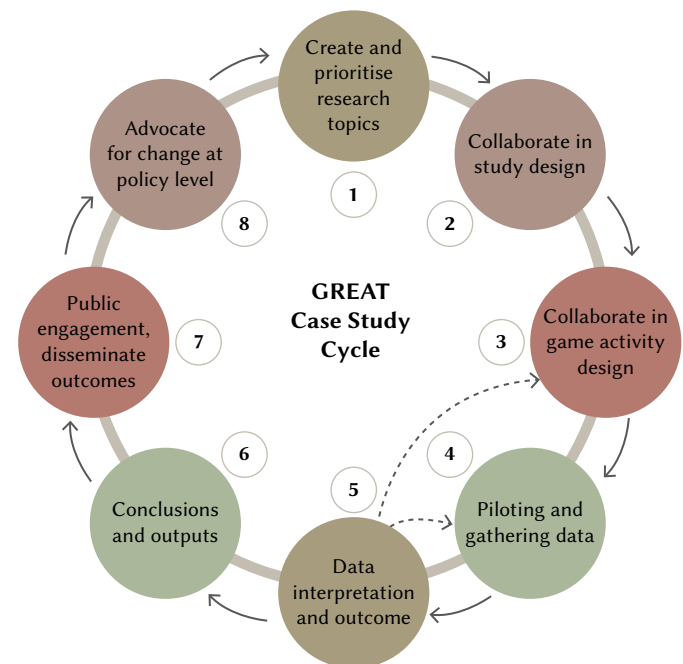

Fig. 1. The GREAT Case Study Cycle.

_____

[1] https://dibl.eu/about-us/

PlanetPlay worked closely with the collaborating studio to design introductory messaging, to encourage as many players as possible to click on an in-game news item. This is important, to avoid inadvertently pre-selecting a certain profile of person with specific opinions on the topic.

B. Tools and Instruments Used

This study makes use of the PlanetPlay system and infrastructure, which is designed to collect insights from video and players of highly popular mobile, computer or console games through a short set of questions that are promoted by game studios at scale to their player base. The PlanetPlay survey system has been developed incrementally since 2020 through a series of practical implementations to meet the information needs of clients.

1. The PlanetPlay System

The PlanetPlay system consists of the survey web application, the infrastructure to host surveys and collect and process data, a 'LiveOps' (live operations) dashboard to monitor activity, and internal tools to assist in building and deploying them. In a typical deployment, players are asked about 7-10 questions centered around a single topic.

After a brief introduction, questions either ask about the respondent's sentiments (opinions on a topic) or knowledge (where there is a clear right answer), as shown in Fig. 2. Questions are usually shown one at a time, with fixed answer options given via buttons. There is no technical limit to the number of alternative answers, but to limit scrolling and maximize comprehensibility there are usually between four and six. Questions can also be marked as multiple choice. A small number of demographic questions are added at the end, usually gathering data about age, gender and education level. The interface design emphasizes simplicity, prioritizing respondents focus on a clearly delimited task, and so to maximize completion rates. Accordingly, additional features, such as branching, external links, and branching according to users' responses, have to date been consciously avoided.
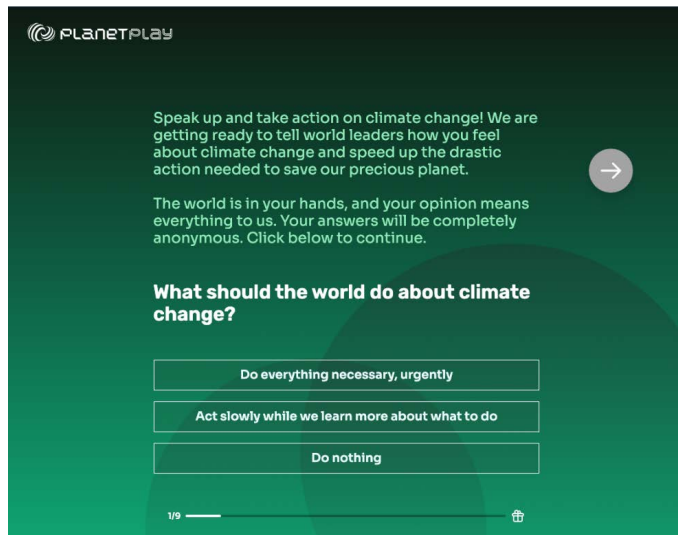
Fig. 2. Global Climate Insights survey introduction and opening question.

2. Infrastructure

The survey system uses a single Next.js app capable of handling multiple sets of survey content, which is hosted on Amazon Web Services via SST[2]. Survey pages are statically generated to improve load times. The preferred language is identified based on web browser

---

[2] https://sst.dev

settings and used if localized content is available. An example survey is available online[3].

Data is collected and previewed with a specialist real-time database platform, Tinybird, that makes it possible to quickly create an infrastructure that scales well. The Tinybird managed Events API receives data directly from the survey web app, continuously aggregating responses to individual questions, so data is not lost if the survey is not completed. To track responses, surveys are tagged with up to 4 identifiers:

- survey: the specific set of content (questions/answers) for a survey.
- source: the name of the game studio and/or their game promoting a survey.
- distribution: one or more specific distribution methods (social media, in-game etc.) used for a survey/source.
- variant: used when A/B testing is employed, e.g. to test response rate changes for UI variations.

A 'LiveOps' dashboard has been developed for the GREAT project using Next.js and the Tremor[4] framework to monitor survey campaigns which form a part of case studies. It is used by PlanetPlay staff and GREAT Project partners too:

- Review recent survey activity and top-level performance indicators for surveys.
- View aggregate answers given to survey questions.
- See geographical and language distribution of responses.
- Trigger data exports (summaries or full data sets).
- List links to the administration panels of GREAT serious games, using the DIBL platform of Serious Games Interactive[5].
- Monitor Tinybird infrastructure resource usage/costs.

The LiveOps dashboard, shown in Fig. 3, also offers a quick way to copy summaries for a particular studio's distribution, so that high level response breakdowns can easily be shared with a game studio without the need for a data analyst.
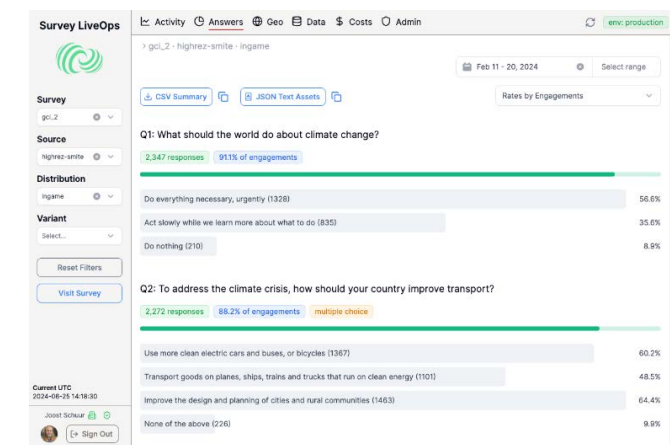
Fig. 3. LiveOps dashboard.

It should be stressed that the LiveOps dashboard is not a tool for data analysis, but rather a quick way of sharing a view of the dataset in a way which is comprehensible to non-technical users. When the survey is complete, all the data is exported in CSV format, using three tables to minimize file size, then processed via the preferred

---

[3] https://survey.planetplay.com/survey/gci_2?source=casestudy&distribution=sampleurl

[4] https://tremor.so

[5] https://www.seriousgames.net/en/

data analysis tools of the analyst, in this case SPSS. Aggregate survey responses can also be made available to the public on the PlanetPlay website in a data panel, as well as being published as open data in line with GREAT project policy.

3. Survey Questions

The questions to be included in the embedded survey were developed in close collaboration with the United Nations Development Programme6 (UNDP), who were the policy stakeholders involved in the study. The final set of questions was as follows:

- What should the world do about climate change?
- To address the climate crisis, how should your country improve transport?
- To address the climate crisis, what should your country do about energy?
- To address the climate crisis, what do you think your country should do about nature?
- To address the climate crisis, what should governments do about farms and food?
- To address the climate crisis, what should governments do about the economy?
- How can your country better protect people from extreme storms, flooding, droughts, forest fires and other climate impacts?
- Do you think games can contribute to resolving climate change?
- In your view How do you think games could best help tackle climate change?
- What climate topics do you think games can best cover?
- Age.
- Gender.
- How old were you when you left education?

C. Participants

Policy stakeholders and other case study partners engage with the GREAT project for a variety of different reasons. The primary purpose of the stakeholders in this case study, UNDP, was to explore methods and approaches that could engage the global community in their Climate change policy discussions, as framed by their organization objectives, which include the engagement of citizens in support of the achievement Nationally Determined Contributions (NDC). Each country which is party to the Paris Agreement [44] is required to establish a NDC plan to adapt to climate impacts. and update it every five years.

PlanetPlay works with game developers and publishers to reach survey respondents from their player base in several ways, including through, paid placement of advertisements, directly in-game, or through social media channels, QR codes and other channels such as a newsletter or homepage link.

The case study explored two approaches to reaching participants through embedded content in the game. The first involved exploring the use of paid placement using the Meta ads platform on Facebook and Instagram. A/B testing was carried out to see if a better response was obtained when showing the first question directly, or when showing an introductory screen.

The second approach was in-game roll out. This access was not paid for but was the result of direct negotiation with Game studios to allow the incorporation of the activity within the game. Prior experience had indicated that in-game promotions implemented in collaboration with a publisher tended to get the highest volume of

responses and response/completion rates. Candidate games usually have a 'live service' model (also known as Games as a Service) and are designed for a long lifespan with a continuous release schedule of new content. To support this model, they usually have existing in-game news/messaging systems that can be leveraged to promote a survey.

In collaboration with UNDP stakeholders, the game studio Hi-Rez was selected, and the embedded QR code is shown in Fig. 4. SMITE, originally published in 2014, is a free-to-play third person Multiplayer Online Battle Arena (MOBA) digital game published on multiple platforms including the Microsoft X Box, Sony PlayStation 4, Nintendo Switch and Amazon Luna. Players control a 'god' 'goddess' or other mythological figures to participate in team-based combat activities with other players and non-player characters (NPC) 'minions'.

Fig. 4. SMITE in-game promotion on a console platform, with Global Climate Insights survey QR code.

The game has multiple modes, supports an active e-sports community, and currently has over ten million global players. This game was chosen as having a high number of users but not so many that data management would be problematic for an exploratory study. The graphics shown in Fig. 5 were used to link to the embedded survey.
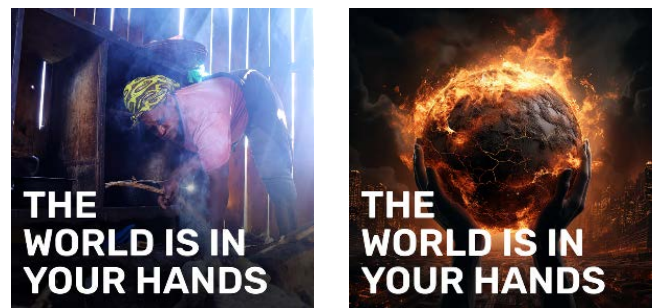
Fig. 5. The graphics used to link to the survey.

Delivery of the survey can be tightly targeted, down to the level of a town or part of a city. However, it was decided to gather data at a global level, to explore the responses across different countries.

D. Procedure

The procedure followed the overall structure of the GREAT case study steps 1 to 6.

Step 1 of the case study cycle was very brief, as UNDP already had a clear view on the topic it wanted to address, i.e. perspectives of citizens on the appropriate actions to address climate change which could inform NDCs.

Corresponding to step 2 of the case study cycle, the project team worked with UNDP and Hi-Rez Studios to design the intervention. It was decided to combine paid placement and in-game roll out, with

---

6 https://www.undp.org/

Hi-Rez studios providing the case study with free access to the game SMITE through a QR code linking to the survey in their main menu.

The design of activities, step 3 of the case study cycle, was limited to the collaborative authoring of questions (see previous section) with the policy stakeholder UNDP.

In step 4, data was collected anonymously, without the respondent needing an account or being tracked or identified via cookies. This streamlined the user experience and allowed a full survey to be completed in under a minute for about 10 questions. Answers were submitted to the survey system when the respondent proceeded to a new question, ensuring that even if a survey instance was not completed, as many answers as possible were collected. Along with the language used, the respondent's country and city were deduced and logged on the basis of the HTTPS headers from Cloudfront, together with the browser's user agent string which includes their device and operating system type. Identification at this level maintains anonymity but offers additional dimensions for segmentation during data analysis.

Efforts were made to discourage multiple responses from the same person, by showing a message reminding the player if they have already participated in a particular survey rather than taking them to the questions. However, this is browser dependent and can be circumnavigated by a technically aware respondent. Additionally, the hashed IP address of the respondent was also logged, and this can be taken into consideration during analysis to flag potential repeated responses. As with the interaction design, this avoidance of all identity management reflects the priority given to a streamlined user experience in order to achieve large scale responses. Finally, the survey/source/distribution/variant identifiers used during the survey promotion were also logged per session.

In addition to the answers given, the dwell time of a question before an answer was captured, giving an indication of how much time someone might have thought about their response. The survey system also tracked 'events', e.g. when a survey is first loaded up, the initial engagement, when questions are shown, and if a link at the end of the survey is followed. This allowed engagement and completion rates to be calculated to measure the effectiveness of a survey instance with a specific partner and distribution method.

Step 5, Data Interpretation and Outcomes, and Step 6, Conclusions and Outputs, are discussed in the following two sections.

V. Results

A. Response Rate by Distribution Method

TABLE I. Response Rates by Distribution Method

	Paid placement (advertisement)	In-Game Roll out
Reach/First page load	7,257	4,352
Engagement	398 (5%)	2,539 (58%)
Completion	179 (45%)	2,148 (84%)
Community Sign Up	18 (10%)	282 (13%)

The categories in Table I, above, are defined as follows:

- *Reach*: people who saw the initial advertisement or survey screen page.
- *Engagement*: people who performed an action on the initial page.
- *Completion*: people who completed the survey.
- *Community sign up*: people who responded to the prompt "Join us and your favorite games, to fight against climate change and save our planet! Track our collective progress and take action now!" by

creating a PlanetPlay account.

Given the poor response rate for paid placement, this data was discarded in further analysis. Not only was it considered that the low level of engagement might be associated with poor quality data, but also this enabled us to focus clearly on evaluating the results of the in-game roll out.

There was a small increase in engagement when an introductory page was shown (58.1% vs 61.6%). Similarly, there was slight preference for the graphic link showing a woman at work rather than a burning planet. However, both effects were small, and are not considered to be statistically relevant.

B. Geographical Distribution

The case study was open to participants from around the world (see Table II). Every time a player visited the survey a 'session' was created, including people who simply visited the introductory page. A session is the parent of all the answer and event data from a specific user in a browser.

TABLE II. Sessions by Country

Country	Sessions
United States	2119
Canada	289
Brazil	225
United Kingdom	184
Mexico	157
Spain	156
Argentina	138
Germany	126
France	114
Russia	93
Colombia	50
TOTAL	3651

From the total of 3651 sessions, 2200 completed responses were obtained.

Response rates can also be tracked by city. As can be seen in Table III, these were widely distributed.

TABLE III. Response Rates by City

City	Country	Sessions
Buenos Aires	Argentina	27
Chicago	United States	27
Lima	Peru	26
Los Angeles	United States	22
Moscow	Russia	21
Bogotá	Colombia	21
Montreal	Canada	21
Houston	United States	19

C. Data Analysis

Data analysis corresponds to Step 5 of the case study cycle. The full data set from the in-game placement has been made available as open data for inspection or further analysis by interested parties [45]. Several interesting trends were identified, as we discuss below. However, as shown in Fig. 6, the data was strongly skewed to the 18-35 age group, and the male gender. In view of the small number of respondents in some age groups, and the even smaller number of female respondents
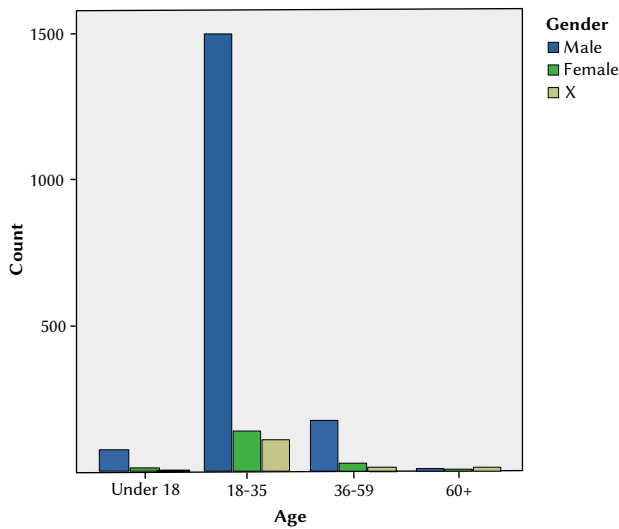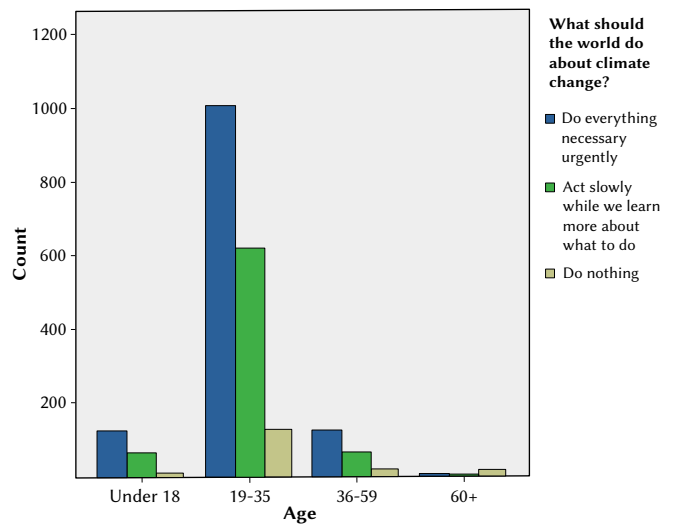
Fig. 6. Age and gender of respondents.

Fig. 7. Age / "What should the world do about climate change?".

within those age groups, the results should be seen as indicative what can be achieved using this approach and should not be generalized to a wider population.

It is not surprising that there should be an imbalance in the gender distribution, but it was not anticipated that this would be so extreme. It is often reported that the male-to-female ratio of game players is close to 50:50, but we obtained far fewer female respondents. We examined the data to see if this could be due to over-representation of console players (a platform that is often seen as being more male dominated) but the response rates were similar for PlayStation (male 85.15%, female 8.25% and other 6.60% (n=303)) and for the data overall (male 83.72%, female 8.78%, other 7.50% (n=2107)). It is therefore concluded that the imbalance reflects an unexpectedly large gender divergence among players of SMITE.

In responses to questions on specific climate change strategies, age was not a significant factor in views on the conservation of forests and land or promotion of plant-based diets. However, the over 60s diverged significantly in their responses for transport, energy, the economy and protecting people.

Gender differences were observed across variables, with statistical differences related to improving transportation of goods (X2 (2, 2107) = 12.263, p = 0.002), design and planning of cities and communities (X2 (2, 2107) = 6.579, p = 0.037), supporting communities (X2 (2, 2107) = 29.186, p = 0.000), using renewable power (X2 (2, 2107) = 7.696, p = 0.021), reducing food waste (X2 (2, 2107) = 10.808, p = 0.004), building infrastructure and conserving nature (X2 (2, 2107) = 15.828, p = 0.000) as well all on all economical interventions (p<0.05). However, there was agreement on several topics, such as using electric vehicles, wasting less energy, stopping fuel burning, conservation of forests and land, and promoting plant-based diets. All the respondents irrespective of gender agreed that games can support climate change initiatives in areas like transport, food and farms, and protecting people.

An interesting result is that views on the appropriate response to climate change vary substantially with age, as shown in Fig. 7. Respondents under 60 were strongly in favor of "Do everything necessary urgently" whereas the majority of those over 60 chose "Do nothing". The relatively small number of respondents over 60 means that the reliability of this result should be treated with caution. It could also be argued that, because of their small numbers, players of SMITE who are over 60 may be a niche population with characteristic attitudes, whereas this is less likely for age groups where the game is more widely played. Consequently, the over 60s may be less typical of

citizens as a whole than are those age groups which are more strongly represented.

Regarding responses to questions on specific climate change strategies, the chi-square results indicated that age was not a significant factor in views on the conservation of forests and land (X2 (3, 2107) = 7.655, p = 0.054) or promotion of plant-based diets (X2 (3, 2107) = 3.075, p = 0.380). However, significant differences were seen in the areas of transport, energy, the economy, protecting people (p<0.05) but not for nature (X2 (3, 2107) = 7.212, p = 0.065) and food and farms (X2 (3, 2107) = 4.976, p = 0.174).

Regarding level of education (see Fig. 8), those respondents who left school after the age of 17 were strongly convinced that games can contribute to solving the climate crisis, whereas those who had left before 16 were largely unsure. Those who had never been to school were predominantly negative or unsure about this topic, but their numbers were too low to be reliable.
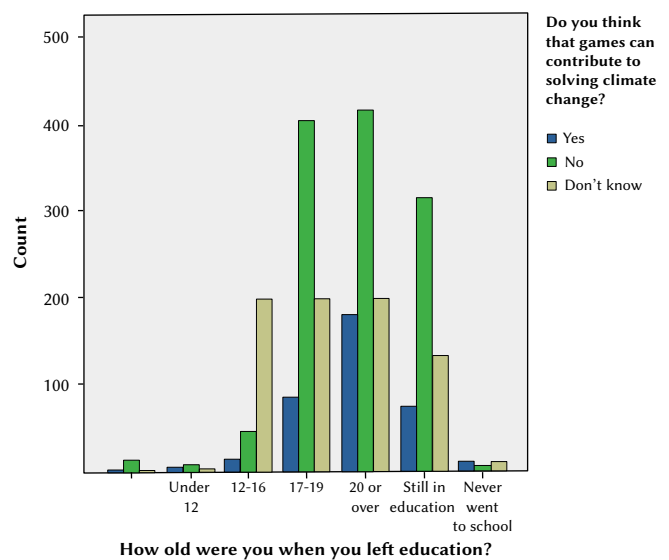

Fig. 8. Educational level of respondents / "Do you think that games can contribute to solving climate change".

The statistical chi-square test results revealed significant differences in views on climate change across different education levels for most variables, as all p-values were below the threshold of 0.05. However, there was no significant difference for variable concerning

conservation of forest and land as indicated by chi-square value ($X2 (6, 2255) = 10.023$, $p = 0.124$) indicating that the respondent views do not vary significantly by education level.

VI. Discussion

We now discuss our results in terms of the research questions set out in Section II.

RQ1 Which Games-based Activities Can Be Used to Elicit, Represent and Communicate Citizens' Views on Policy Dilemmas?

The results show that embedding questions within a popular commercial game is a viable strategy for gathering citizens' views on policy dilemmas, and, in particular, with regards to climate change. The data collection, data formats and procedures tested in this case study proved to be fit for purpose. It is concluded that the approach does enable a link between policy makers and the targeted sector of games culture, but concerns are raised about potential distortion, which need to be addressed (see RQ2, below).

More specifically, the results show that in-game roll out in collaboration with a game studio is far more effective than paid placement, as shown in Table I. One factor which may influence this result is that in-game roll out enables a respondent to scan a QR code on their mobile phone, and to answer the questions there while leaving the state of the game unchanged on their PC or console.

Our survey was delivered through a QR code link on the main menu of the game. Other games may have a news system that shows regular updates and a survey can be promoted there, with a link button opening an external browser window. These existing news (or player messaging) systems are primarily designed to promote new game content and events, which keep players engaged and eventually monetized, so they could also be a very effective and visible way to reach players with a survey.

Our experience of collaboration suggests two motivations which explain why game studios might be willing to offer access to organizations such as UNDP. Firstly, this shows that a studio is using its power as a media provider in a socially responsible way, contributes to meeting their Corporate and Social Responsibility commitments, and improves their image and brand association. Secondly, it helps the studio to better understand the issues that are important to their player base, which assists them in game design and content decisions.

RQ2 How Effective Are Games-based Activities in Eliciting, Representing and Communicating Citizens' Views on Policy Dilemmas?

The results show that our approach is an effective alternative to gamified surveys. Three challenges to be addressed were identified in Section III above. Regarding challenge one 'engagement levels', the method and tools were effective in obtaining a large sample size with relatively little technical effort, once agreement had been reached with the studio. The response rate was good, as was the quality of the engagement, particularly in the in-game roll out, where 84% completed the survey once they had started, and 13% took the additional step of enrolling in the PlanetPlay climate change community. We conclude that the approach is effective in terms of engagement.

Challenge two was 'measurement error' caused by the gamified elements of a gamified survey. Our separation of the survey from the mechanics and interactions of the host game worked well, and while we did not directly evaluate comparative measurement error, this strategy addresses some of the underlying concerns of this critique. However, the very large skewing of the data towards males between 18 and 25 years old is a concern, as it may misrepresent the views of all citizens. Information about the profile of players of particular games is published for different genres [46] [47], but the skew in our data is much more substantial than we anticipated from such high level analyses of player profiles.

RQ3 How Efficient Is the Use of Games-based Activities in Eliciting, Representing and Communicating Citizens' Views on Policy Dilemmas?

Challenge three identified in Section III was 'effort', which relates strongly to our RQ3. In this respect we can offer some encouraging initial results. The PlanetPlay infrastructure enabled the intervention to be designed and delivered and the data managed with a relatively low level of skilled technical input, in the order of person days rather than person months. The system was easily able to handle the number of respondents, and there were no indications that the system would not be scalable to very large numbers. Our experience in this case study, therefore, was that the system used is highly efficient in technical terms. This does not take into consideration the very considerable effort involved in creating the system, which would need to be replicated by anyone adopting the approach who did not partner with PlanetPlay.

As discussed in relation to RQ2, in-game rollout was a far more effective way of engaging with gamers than paid placement. However, the case study underlined the essential role played by the GREAT project, and specifically partner PlanetPlay, in mediating between the policy stakeholder UNDP and the games industry. This effort required to establish collaboration with games studios to obtain access to their platforms also needs to be taken into consideration when assessing the efficiency of the approach. However, this is hard to quantify, as it varies greatly from case to case, and depends strongly on the strength of the existing connections of the team carrying out the case study with appropriate sections of the games industry.

VII. Reflections and Future Work

A. Limitations

An important limitation of this study is the highly skewed age and gender of the respondents. This constrains the generalizability of interesting results, such as the preference of over 60s for doing nothing to address climate change, in contrast to the views of younger people.

In terms of the overall GREAT case study methodology, this study is limited by focusing only on the first five steps of the cycle. It therefore does not consider the value to stakeholders of the information obtained, nor its use to inform their decision making.

For both of the above reasons, the value of the study lies more in the validation of the method, rather than in the impact or value of the specific data which was generated about citizens views.

It should also be recognized that our approach corresponds to the second of Mayer's three categories cited in Section III, e-engaging, and therefore involves "top-down consultation of citizens by government or parliament". This is not presented as an alternative to e-empowering, but rather as the provision of a tool for linking citizens and policymakers at a scale which would not be possible for approaches which involve games players in collaborative policy making.

B. Implications

This case study has demonstrated that the in-game roll out approach to obtaining the views of citizens is effective and can generate information which is of value to policy stakeholders. This provides evidence in support of the use of the approach by policymakers and other policy stakeholders who experience difficulties in establishing a full picture of the views of citizens with regard to policy dilemmas.

If the policy stakeholder is interested in the views of citizens over large geographic areas, then in-game roll out is appropriate, as the intervention is delivered in the regions where a game is marketed,

and indeed this is often global. In collaboration with studios, an in-game roll out study can select games for the study with contrasting age, gender and geographic profiles. For example, Table II shows that SMITE would be a particularly good choice for an inquiry centered on the United States. Alternatively, if policy stakeholders have a need to work with a highly focused population, paid placement of playable advertisements is more appropriate, as this can be restricted to a particular city or region. Very specific targeting can be achieved by making use of more detailed data from a paid service provided by companies such as Ironsource and Loopme. Consequently, paid placement may still be a valuable option for more local inquiries, despite the lower response rate we have reported.

The work reported here may be seen as a case study which examines the potential of maximizing scale and ease of participation at the cost of the depth of games players' engagement or collaboration. This strategy has been shown to be effective, in as much as it has been well received by the policy stakeholder, UNDP, who have committed to a larger scale study informed by the results and limitations of this case study, being implemented at the time of writing. However, other approaches, with different trade-offs between these aspects, would potentially be equally effective. These would result in different benefits and costs, which may be more suitable for other contexts. Indeed, in parallel with the work reported here, the GREAT project is working with serious games to explore citizens' views on policy dilemmas in intensive interactions between small numbers of participants.

In future studies, greater attention needs to be paid to the selection of the game or games which host the survey, in order to achieve a better balance of gender and age. This can be done through selection of appropriate games, with the possibility of adapting the distribution of the survey in the light of data collected in the 'LiveOps' dashboard. This approach will not guarantee equal numbers of respondents from all genders and age groups but should provide sufficient responses from all genders to ensure the validity of the results.

C. The Potential Use of Artificial Intelligence

A possible strategy which could change the equation between the scale of an intervention and the depth of the interactions would be to make use of artificial intelligence (AI). In raising this possibility we are not referring to the automated generation of surveys or survey questions, as proposed by Gonzalez Bonorino [48] and by numerous websites such as responsly.com. Rather we see two potential techniques. Firstly, as suggested by Xiao et al. [49], AI-based chatbots could be used to generate interactions with players. This could act as a virtual equivalent of the familiar researcher with a clipboard who engages with a respondent. While one might associate chatbots with text, perhaps the most compelling application would be to generate conversational spoken language interactions between an app and a game player, which would provide richer data about players views than is available with text. The interactions could be as short or extended as the designer wished or could adapt to the respondent's input. To address privacy concerns, it would be advisable to transform the speech into text, and to store the results of sentiment analysis, in as close to real-time as possible, and to avoid storing a recording of the player's voice. Such a use of chatbots could, in principle, be compatible with the broad approach that we have discussed.

Secondly, it would be possible to implement AI based non-player characters which could interact with players of a game. Johnson's description of the goal of AI in commercial video games, from 2014, remains relevant:

> ...to produce believable behavior that is predictable and unpredictable and feels as if the player is being challenged and given interesting decisions to make in relation to an intelligent agent or character. [50]

It is easy to see how giving players "interesting decisions to make"

in relation to policy dilemmas could provide rich data about players views. Such an approach moves away from the highly focused large-scale interactions in the present case study. It would also require substantial investment for each game in which it was included and would engage players in extended interactions. Moreover, it is doubtful that games studios would generally be willing to provide the access to their games which would be needed to implement this kind of interaction, as it would interfere directly in the gameplay. Consequently, it seems more likely that this approach would require the creation of a specific project to create a purpose-built game, with an alliance between researchers and industry.

REFERENCES

[1] R. S. Contreras-Espinosa and A. Blanco-M, "A Literature Review of E-government Services with Gamification Elements", *International Journal of Public Administration*, vol. 45, no. 13, pp. 964–980, 2021, doi: 10.1080/01900692.2021.1930042.

[2] R. S. Foa, A. Klassen, M. Slade, A. Rand, and R. Collins, "The Global Satisfaction with Democracy Report 2020", Bennett Institute for Public Policy, Cambridge, UK, 2020. Accessed: Aug. 08, 2024. [Online]. Available: https://www.cam.ac.uk/system/files/report2020_003.pdf

[3] United Nations, "Trust in public institutions: Trends and implications for economic security (Policy Brief 108)", United Nations Department of Economic and Social Affairs, 2021. Accessed: Aug. 08, 2024. [Online]. Available: https://www.un.org/development/desa/dpad/wp-content/uploads/sites/45/publication/PB_108.pdf

[4] D. Freeman *et al.*, "Coronavirus Conspiracy Beliefs, Mistrust, and Compliance with Government Guidelines in England", *Psychological Medicine*, vol. 52, pp. 251–263, 2020, doi: 10.1017/S0033291720001890.

[5] A. Caplanova, R. Sivak, and E. Szakadatova, "Institutional Trust and Compliance with Measures to Fight COVID-19", *International Advances in Economic Research*, vol. 27, no. 1, pp. 47–60, 2021, doi: 10.1007/s11294-021-09818-3.

[6] M. Morelli, "Sad populism and the policies of hope", *EconPol Forum,* ISSN 2752-1184, CESifo GmbH, Munich, vol. 25, no. 2, pp. 40-42.

[7] J. Visser, "Trade Unions in the Balance, ILO ACTRAV Working Paper", 2019, Accessed: Aug. 05, 2023. [Online]. Available: https://etufegypt.com/wp-content/uploads/2019/10/wcms_722482.pdf

[8] R. F. Inglehart, *Religion's Sudden Decline.* Oxford University Press, UK, 2021. doi: https://doi.org/10.1093/oso/9780197547045.001.0001

[9] [N. Forman-Katz, "Americans are following the news less closely than they used to", Pew Research Center, Washington DC, USA. Accessed: Jun. 14, 2024. [Online]. Available: https://www.pewresearch.org/short-reads/2023/10/24/americans-are-following-the-news-less-closely-than-they-used-to/

[10] Pew Research Center, 'News Platform Fact Sheet'. Accessed: Jun. 14, 2024. [Online]. Available: https://www.pewresearch.org/journalism/fact-sheet/news-platform-fact-sheet/

[11] J. A. Lischka, "A Badge of Honor?: How *The New York Times* discredits President Trump's fake news accusations", *Journalism Studies*, vol. 20, no.

2, pp. 287–304, Jan. 2019, doi: 10.1080/1461670X.2017.1375385.

[12] S. Kreps, R. M. McCain, and 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, doi: 10.1017/XPS.2020.37.

[13] Future Markets Insights, "Video Game Market". Accessed: Jul. 04, 2024. [Online]. Available: https://www.futuremarketinsights.com/reports/video-game-market

[14] A. Shaw, "What Is Video Game Culture? Cultural Studies and Game Studies", *Games and Culture*, vol. 5, no. 4, pp. 403–424, Oct. 2010, doi: 10.1177/1555412009360414.

[15] F. Lunning, *Cosplay: The Fictional Mode of Existence*. University of Minnesota Press, USA, 2022. ISBN: 1452967466, 9781452967462

[16] M. Consalvo, *Cheating: Gaining Advantage in Videogames*. Cambridge MA: MIT Press, 2007, ISBN: 0262033658, 9780262033657.

[17] P. Vergel, D. La parra-Casado, and C. Vives-Cases, "Examining Cybersexism in Online Gaming Communities: A Scoping Review", *Trauma, Violence, & Abuse*, vol. 25, no. 2, pp. 1201–1218, Apr. 2024, doi: 10.1177/15248380231176059.

[18] European Games Developer Association, "All About Video Games: European Key Facts 2022", 2023. Accessed: Jun. 14, 2024 [Online]. Available: https://www.videogameseurope.eu/wp-content/uploads/2023/08/Video-Games-Europe_Key-Facts-2022_FINAL.pdf

[19] A. Drummond, J. D. Sauer, and C. J. Ferguson, "Do longitudinal studies support long-term relationships between aggressive game play and youth aggressive behaviour? A meta-analytic examination", *Royal Society Open Science.*, vol. 7, no. 7, p. 200373, Jul. 2020, doi: 10.1098/rsos.200373.

[20] J. Burkhardt and W. Lenhard, "A Meta-Analysis on the Longitudinal, Age-Dependent Effects of Violent Video Games on Aggression", *Media Psychology*, vol. 25, no. 3, pp. 499–512, Sep. 2021, doi: 10.1080/15213269.2021.1980729.

[21] V. Rideout, A. Peebles, S. Mann, and M. B. Robb, "The Common Sense Census: Media Use by Tweens and Teens, 2021", Common Sense, San Francisco, CA., USA, 2022. Accessed: Jun. 16, 2024. [Online]. Available: https://www.commonsensemedia.org/sites/default/files/research/report/8-18-census-integrated-report-final-web_0.pdf

[22] M. L. Hanin, "Theorizing Digital Distraction", *Philosophy and Technology*, vol. 34, no. 2, pp. 395–406, Jun. 2021, doi: 10.1007/s13347-020-00394-8.

[23] J. Williams, *Stand out of our light: Freedom and resistance in the attention economy*. Cambridge University Press, Cambridge, UK, 2018.

[24] J. L. Kröger, P. Raschke, J. Percy Campbell, and S. Ullrich, "Surveilling the gamers: Privacy impacts of the video game industry", *Entertainment Computing*, vol. 44, p. 100537, Jan. 2023, doi: 10.1016/j.entcom.2022.100537.

[25] Council of the European Union, "Council conclusions on enhancing the cultural and creative dimension of the European video games sector", Council of the European Union, Brussels, 15901/23, 2023. Accessed: Jun. 10, 2024. [Online]. Available: https://data.consilium.europa.eu/doc/document/ST-15901-2023-INIT/en/pdf

[26] Q. Yu, K. Yu, and B. Li, "Can gamification enhance online learning? Evidence from a meta-analysis", *Education and Information Technologies*, vol. 29, no. 4, pp. 4055–4083, Mar. 2024, doi: 10.1007/s10639-023-11977-1.

[27] X. Guan, C. Sun, G.-J. Hwang, K. Xue, and J. Wang, "Applying game-based learning in primary education: a systematic review of journal publications from 2010 to 2020", *Interactive Learning Environments*, vol. 32, pp. 1–23, Jul. 2022, doi: 10.1080/10494820.2022.2091611.

[28] D. Vlachopoulos and A. Makri, "The effect of games and simulations on higher education: a systematic literature review", *International Journal of Educational Technology in Higher Education*, vol. 14, no. 1, pp. 14–22, Dec. 2017, doi: 10.1186/s41239-017-0062-1.

[29] S. Janakiraman, S. Watson, and W. Watson, "Using Game-based Learning to Facilitate Attitude Change for Environmental Sustainability", *Journal of Education for Sustainable Development*, vol. 12, pp. 176–185, Sep. 2018, doi: 10.1177/0973408218783286.

[30] D. B. Dhiman, "Games as Tools for Social Change Communication: A Critical Review", *Global Media Journal*, vol. 21, no. 61, pp. 137–141, 2023, doi: 10.36648/1550-7521.21.61.357.

[31] K. M. Kapp, *The Gamification of Learning and Instruction: Game-based Methods and Strategies for Training and Education*. John Wiley & Sons, USA, 2012.

[32] F. Keusch and C. Zhang, "A Review of Issues in Gamified Surveys", *Social Science Computer Review*, vol. 35, no. 2, pp. 147–166, Apr. 2017, doi: 10.1177/0894439315608451.

[33] J. Hamari, J. Koivisto, and H. Sarsa, "Does Gamification Work? -- A Literature Review of Empirical Studies on Gamification", in *2014 47th Hawaii International Conference on System Sciences*, Waikoloa, HI: IEEE, Jan. 2014, pp. 3025–3034. doi: 10.1109/HICSS.2014.377.

[34] N. Yee, "The Demographics, Motivations and Derived Experiences of Users of Massively Multi-User Online Graphical Environments", *Presence: Teleoperators and Virtual Environments*, 15, pp. 309-329, 2006. doi: 10.1162/pres.15.3.309.

[35] A. Blanco-M, R. S. Contreras-Espinosa, and J. Solé-Casals, "Clustering Users to Determine the Most Suitable Gamification Elements", *Sensors*, vol. 22, no. 1, p. 308, Dec. 2021, doi: 10.3390/s22010308.

[36] S. R. Arnstein, "A Ladder of Citizen Participation", *American Institute of Planners Journal*, vol. 35, no. 4, pp. 216–224, 1969, doi: 10.1080/01944366908977225

[37] I. S. Mayer, "The Gaming of Policy and the Politics of Gaming: A Review", *Simulation & Gaming*, vol. 40, no. 6, pp. 825–862, Dec. 2009, doi: 10.1177/1046878109346456.

[38] S.-K. Thiel, M. Reisinger, K. Röderer, and P. Fröhlich, "Playing (with) Democracy: A Review of Gamified Participation Approaches", *eJournal of eDemocracy and Open Government*, vol. 8, no. 3, pp. 32–60, Dec. 2016, doi: 10.29379/jedem.v8i3.440.

[39] J. Gastil and M. Broghammer, "Linking Theories of Motivation, Game Mechanics, and Public Deliberation to Design an Online System for Participatory Budgeting", *Political Studies*, vol. 69, no. 1, pp. 7–25, Feb. 2021, doi: 10.1177/0032321719890815.

[40] K. W. R. Oliveira and M. M. V. Paula, "Gamification of Online Surveys: A Systematic Mapping", *IEEE Transactions on Games*, vol. 13, no. 3, pp. 300–309, Sep. 2021, doi: 10.1109/TG.2020.3004366.

[41] J. Harms, D. Seitz, C. Wimmer, K. Kappel, and T. Grechenig, "Low-Cost Gamification of Online Surveys: Improving the User Experience through Achievement Badges", in *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, London United Kingdom: ACM, Oct. 2015, pp. 109–113. doi: 10.1145/2793107.2793146.

[42] P. Hollins, J. Ower, D. Griffiths, B. Keislinger, K. Koller, and C. Fabian, 'GREAT D4.2 GREAT Case Study Plan', 2024. [Online]. Available: https://doi.org/10.5281/zenodo.12659432

[43] K. Koller, B. Kieslinger, and C. Fabian, 'GREAT D5.1 Evaluation Framework', 2024. Accessed: Jul. 04, 2024. [Online]. Available: DOI 10.5281/zenodo.12637086

[44] UNFCCC, *Paris Agreement*. 2015. Accessed: Jul. 05, 2024. [Online]. Available: https://unfccc.int/files/meetings/paris_nov_2015/application/pdf/paris_agreement_english_.pdf

[45] J. Schuur, J. Ower, A. Garg, P. Hewage, P. Hollins, and D. Griffiths, "Data Set from GREAT Case Study 1 (1.0.0)". Accessed Aug. 05, 2024. [Online] Zenodo. https://doi.org/10.5281/zenodo.12686861, 2024.

[46] N. Yee, "Beyond 50/50: Breaking Down The Percentage of Female Gamers By Genre", Quantic Foundry. Accessed: Jul. 07, 2024. [Online]. Accessed Jul. 05, 2024. [Online]. Available: https://quanticfoundry.com/2017/01/19/female-gamers-by-genre/

[47] GameTree, "Game Demographics By Genre And Platforms". Accessed: Jul. 07, 2024. [Online]. Available: https://gametree.me/blog/global-gamer-insights-report/

[48] A. Gonzalez Bonorino, "Smart Surveys: An Automatic Survey Generation and Analysis Tool", in *Proceedings of the 15th International Conference on Computer Supported Education*, Prague, Czech Republic: Science and Technology Publications, 2023, pp. 113–119. doi: 10.5220/0011985400003470.

[49] Z. Xiao *et al.*, "Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-ended Questions", *ACM Transactions on Computer-Human Interaction.*, vol. 27, no. 3, pp. 1–37, Jun. 2020, doi: 10.1145/3381804.

[50] R. Johnson, "Artificial Intelligence", in *The Routledge Companion to Video Game Studies*, M. J. P. Wolf and B. Perron, Eds., Routledge, NY, USA, 2014, pp. 13–18.

Dai Griffiths

Dai Griffiths, also known as David, has a background in the arts, and holds a PhD from Universitat Pompeu Fabra. He spent the first part of his career working as a teacher in primary, secondary and higher education, as well as in interpersonal skills training in industry, before becoming fascinated by the potential of computers in education. For the past twenty-five years he has worked in the development of educational applications, and as an educational technology researcher, and has published extensively. In this work he became deeply engaged with the tradition of cybernetics. He was appointed Professor at the Institute for Educational Cybernetics at the University of Bolton, where he worked with the Centre for Education Technology Interoperability and Standards (Cetis). He then took on a role in the Department of Education of Bolton University, leading the Department's PhD and Doctor of Education programs. Dai Griffiths is currently a Senior Researcher at the Research Institute for Innovation & Technology in Education (UNIR iTED), Universidad Internacional de La Rioja (UNIR). ORCID: 0000-0002-6863-2456.

Jude Ower

Jude Ower is the founder of Playmob (acquired by Sphaira Innovation / PlanetPlay), an insights platform gathering public sentiment, through games, with the aim to help organizations achieve the SDGs by 2030. Playmob enabled the UN to collect the world's largest data set on climate attitudes, data which is being used in 60+ countries for climate policy decisions, and has also been used for the G20 and IPCC report in 2022. She is also the Co-Founder of the Playing for the Planet Alliance along with UN Environment. The alliance is a group of 49 forward thinking games studios and publishers, such as Supercell, Rovio, Ubisoft, Sybo, Niantic, Xbox and Playstation, with a collective reach of 1.4 billion monthly players. All members have pledged to lower their emissions to become carbon neutral, or better, negative, and to put green nudges into their games to encourage players to take positive climate actions. She has co-written 'Gaming for Good' with Mathias Norvig (CEO of SYBO, creators of Subway Surfer) which launched in February 2024 and discusses how games have the power to solve some of the world's greatest issues, in particular, Climate Change.

Paul Hollins

Paul is a professor of cultural research and Knowledge Exchange at the University of Bolton in the UK and is a Fellow of the Cybernetic Society. His PhD was focused on exploring the efficacy of immersive environments in formal educational settings, awarded by the University of Bolton in the UK. He also has postgraduate degrees in Education, a Master of Science in learning and teaching with Information & communication Technologies awarded by Leeds Becket University in the UK and a Master of Business Administration awarded by Leeds Business School. His first degree was in Management studies. Paul's research interests are varied but lie in the intersection of technology, learning and digital Games. He has also published extensively in Music. He has directed and participated in several research projects in these spaces over the last three decades including the EU framework funded Learning Interoperability Framework Europe (LIFE), Ten Competence and Realising Applied Games Ecosystem (RAGE) and is currently leading the University of Bolton's participation in the EU Horizon funded Games Realising Effective & Affective Transformation (GREAT) research and innovation project. Paul has published over one hundred and thirty academic outputs and has acted as external examiner for several institutions.

Anchal Garg

Dr. Anchal Garg is a Senior Lecturer (Computing) at the University of Bolton. She holds a PhD in Computer Science & Engineering and has over 22 years of teaching experience. Her research interests include learning analytics and artificial intelligence, and she has published several papers in international conferences and journals. She is a Senior Member of IEEE and a member of ACM, IET and CSAB. Along with teaching and research, she is also involved in industry collaboration and knowledge exchange activities and is an accreditor with ABET and IET.

Towards Promoting the Culture of Sharing: Using Blockchain and Artificial Intelligence in an Open Science Platform

Mouna Denden*, Mourad Abed

Univ. Polytechnique Hauts-de-France, LAMIH, CNRS, UMR 8201, 59313, Valenciennes (France)
INSA Hauts-de-France, F-59313 Valenciennes (France)

* Corresponding author: mouna.denden91@gmail.com

uniR
LA UNIVERSIDAD
EN INTERNET

Abstract

Several studies in the literature have proposed the use of artificial intelligence (AI) tools to manage big data and further enhance collaboration between researchers on open science platforms, hence promoting the culture of safely sharing reliable data. Moreover, some other studies further proposed the use of blockchain technology to secure data, provide transparency in data analysis, and also keep track of all collaborations within open science platforms. Despite the importance of AI and blockchain technology in open science platforms, no study, to the best of our knowledge, has implemented and discussed the benefits of using both technologies together or how blockchain can enhance AI systems in open science. Therefore, to address this research gap, this study presents a newly developed open science platform that harnesses the power of AI and blockchain technologies to promote and foster a culture of sharing and seamless collaboration among universities worldwide. This platform was then validated through focus group analysis from the European University for Customised Education (EUNICE) partners, which is the project context of this present study. The findings revealed that the use of AI and blockchain enabled researchers and institutions to share open science more effectively. Specifically, the use of AI features in Open REUNICE enhanced data management processes, particularly by improving metadata accuracy, searchability and reusability, thereby addressing critical needs in research workflows. Additionally, the use of Blockchain was found to play a critical role in addressing legal challenges and enhancing user trust.

I. Introduction

OPEN science is a movement that aims to make scientific research and data accessible to everyone, regardless of their background or affiliation. It involves the sharing of research data, software, and methods, as well as the promotion of transparency and collaboration in scientific research [1]. It has gained significant momentum in recent years as a powerful approach to enhance the quality, efficiency, and impact of scientific research. Consequently, numerous universities worldwide have incorporated open science into their strategic plans to elevate the quality and influence of their research efforts [2]. The open sharing of knowledge has thus become central to the academic process. However, fostering a culture of data sharing requires cooperation between data generators and data users. Moreover, effective sharing requires the establishment of protocols to maintain a comprehensive data history and address the concerns of data generators regarding recognition for their rigorous work, such as authorship attribution [3].

Thus, the provision of an effective open science infrastructure, capable of addressing these concerns comprehensively, is paramount.

A review on open science conducted by Leible et al. [4] identified a list of specific requirements that should be fulfilled in any open science infrastructure, namely: a collaborative environment, open data, open access, no censorship, identity and reputation management, and an extensive system. Integrating these elements is crucial for building an effective and sustainable open science framework that supports and enhances the collaborative nature of modern scientific research.

On the other hand, several challenges related to the implementation of open science practices have been identified in the literature. For instance, open science generates a large volume of data from various stakeholders, including researchers, non-academic actors, and citizens, leading to potential issues with data duplication and the validity of published data. According to Fecher and Friesike [5], managing the quality and trustworthiness of open data remains a significant challenge. Similarly, Borgman [6] highlights the complexities involved

in curating and validating the vast amounts of data produced in open science initiatives. Additionally, the involvement of multiple stakeholders can complicate the data curation process, as discussed by Tenopir et al. [7], who emphasize the need for robust data management practices to ensure the reliability of open science outputs.

Furthermore, several data sharing problems have been identified in the literature, specifically regarding the sharing of personal data, such as medical information and personality-related data [8]. One of the main objectives of open science is to improve collaboration between academic and non-academic stakeholders; however, there is often a lack of traceability regarding the contributions of collaborators [3]. This challenge of traceability is also identified for data analysis, where ensuring the accuracy and provenance of data remains a significant concern.

To improve open science practices, several studies in the literature highlighted the potential of AI and blockchain technologies. For instance, AI can make science more open and accessible by improving data management and facilitating more efficient research workflows [9]. Specifically, AI can be used to analyse large datasets, predict outcomes, identify patterns, and automate repetitive tasks. AI can facilitate interdisciplinary collaborations between scientists from different fields, leading to novel insights and discoveries that would otherwise have been difficult to achieve. Furthermore, generative AI using large language models (LLM) can help scientists generate new research questions and hypotheses [10].

Blockchain, on the other hand, was proposed to track the traceability of researchers' contributions in collaborative projects, thereby ensuring transparency and accountability [11]. Blockchain was also proposed to overcome AI related challenges, such as interpretability issues in AI models [12]. However, despite the importance of blockchain and AI, no study, to the best of our knowledge, has implemented and discussed the benefits of combining these two technologies in one online platform, including open science platforms [13]. Accordingly, it is hypothesised that the combination of AI and blockchain would enable researchers and institutions to share open science more effectively. To investigate this hypothesis, the present study presents a newly developed open science platform, namely OPEN REUNICE, that harnesses the power of AI and blockchain technologies to promote and foster a culture of sharing and seamless collaboration among researchers. Specifically, this study aims to answer the following research questions (RQs):

RQ1. How users perceive the developed features in the OPEN REUNICE platform?

RQ2. How can the OPEN REUNICE platform address open science challenges, such as data management issues and legal obstacles in the literature?

The rest of the paper is organized as follows. Section II discusses the background of open science and its limitations, AI and blockchain technology. Section III presents the description of the implemented open science platform. Section IV presents the methodology of the study. Section V presents and discusses the results. Finally, section VI concludes the paper.

II. Literature Review

A. Open Science

Four categories of activities that rely on open science were proposed by the United Nations Educational, Scientific and Cultural Organization (UNESCO) [1], namely: (1) open science knowledge, which refers to open access to scientific publications, open research data, open educational resources, open source software and source code, and hardware that are available in the public domain or under copyright and licensed under an open license; (2) open science infrastructures which refers to virtual and physical infrastructures used for sharing data, collaboration and digital research services; (3) open engagement of societal actors refers to extended collaboration between scientists and members of the public outside of the scientific community by applying citizen and participatory science, crowdsourcing, crowdfunding, and scientific volunteering; and, (4) open dialogue with other knowledge systems, which refers to the dialogue between different knowledge holders, such as indigenous peoples, marginalised scholars, local communities. Each of these categories focuses on a different aspect of open science, with the goal of promoting the values of openness, transparency, and collaboration in scientific research. Open science is increasingly associated with FAIR (Findable, Accessible, Interoperable, and Reusable) principles to ensure that science resources have the necessary metadata to make them findable, accessible, interoperable, and reusable [14].

Several problems related to the implementation of open science, specifically with data sharing were identified in the literature. For instance, Zhou [15] found that Lack of trust, lack of a conducive knowledge sharing culture, Lack of strong knowledge sharing leadership, and cultural affinity for autonomy are from the main obstacles of data sharing. Zuiderwijk et al. [16] argued that an online infrastructure to support the provision and use of open data must include specific features, which can be categorized as follows: (1) Data Provision which includes the acquisition of data and metadata, along with their validation and enhancement; (2) Data Retrieval and Use which includes the display of data and its retrieval through query mechanisms; (3) Data Linking which refers to the establishment of connections between datasets using metadata, which can be done automatically or manually; (4) User Rating which refers to rating users' contributions on the platform as well as their publications; and, (5) User Cooperation which refers to understanding users' preferences, responsibilities, and behaviours. Specifically, some of these features can be achieved using AI, such as data linking to find similar works and user cooperation to understand users' behaviours [17]. Consequently, the subsequent section will discuss how AI can be leveraged to overcome certain limitations of open science, enhancing collaboration and data integration.

B. Artificial Intelligence

AI was broadly described by Baker and Smith [18] as "computers which perform cognitive tasks, usually associated with human minds, particularly learning and problem-solving" (p.10). According to this definition, AI encompasses a wide range of tools and techniques, including machine learning, data mining, neural networks, natural language processing (NLP), and numerous algorithms.

AI plays a pivotal role in revolutionizing open science by offering innovative solutions to address the challenges associated with data management, analysis, and collaboration. Notably, machine learning tools, data visualization, and intelligent decision-making are among the most commonly used AI techniques in open science [19]. For instance, machine learning algorithms have been instrumental in automating data processing tasks, enabling researchers to sift through vast datasets efficiently. Moreover, AI-driven tools contribute to enhancing reproducibility by automating experimental workflows and aiding in the validation of scientific findings [20]. The utilization of AI fosters greater transparency and accessibility, allowing researchers to share data, methodologies, and results seamlessly. Gundersen et al. [20] also claimed that the use of AI to generate sufficient documentation for publications, such as basic metadata (title, abstract, keywords, etc.) and digital object identifier (DOI) or URL to ensure permanent accessibility, can facilitate reproducibility in research. Similarly, Patra et al. [17] found that AI can help in generating accurate metadata,

which enhances the effectiveness of dataset recommendations in open science platform. A study by Olivetti et al. [21] explores the application of NLP in scientific literature demonstrating how AI tools can facilitate the extraction and synthesis of information from a vast corpus of scholarly articles, further promoting open access to knowledge. Furthermore, Bail [22] discussed the use of generative AI to identify novel research questions, since it was capable of providing a broad perspective on many different scientific fields. These advancements underscore the transformative impact of AI on open science, paving the way for collaborative and data-driven research practices.

The integration of AI into open science, while promising, confronts a range of challenges as discussed in the existing literature. For instance, ethical considerations regarding the responsible use of AI, particularly in handling sensitive data, were highlighted by Mittelstadt et al. [23] who emphasize the importance of addressing ethical concerns to maintain trust in AI applications. Interpretability issues in AI models, commonly referred to as the "black-box" problem, are articulated by many researchers, underscoring the need for transparent and explainable AI systems in open science [24] [25] [26]. Interoperability challenges are discussed by Mons et al. [27], emphasizing the necessity for standardized data formats and tools to facilitate seamless collaboration across diverse research environments. The demand for substantial computational resources and expertise is acknowledged by Amodei et al. [28], indicating potential barriers to the widespread adoption of AI in open science. Furthermore, security attacks on data and learning algorithms in AI based open science platforms were highlighted by Shinde et al. [29]. In order to overcome the identified AI challenges, several studies in the literature proposed the use of blockchain technology. The next subsequent section presents the blockchain technology and its benefits and challenges.

C. Blockchain

Blockchain is a secured distributed ledger technology that ensures the security and reliability of transactions by providing a digital record of every transaction that has ever taken place on the network [30] [11]. Blockchain organizes its data into blocks that are cryptographically and chronologically linked together. Additionally, it utilizes various consensus mechanisms and smart contracts [31].

Blockchain has the potential to offer numerous benefits across different fields and applications due to its inherent characteristics, namely decentralization, transparency, immutability, better security, anonymity, cost reduction, and autonomy [24]. In the context of open science, Leible et al. [4] demonstrated that blockchain technology serves as a suitable infrastructure for open science, as the characteristics of blockchain align well with the requirements of open science. For instance, blockchain technology can be used to address reproducibility of findings in published articles and experiments issues [32]. The immutability, append-only functionality, and a transparent log of all transactions inherent in blockchain can offer visibility to all users, ensuring transparency across every step within a system. Additionally, the decentralization empowers researchers to construct their individual open ecosystem for research data, metadata, and communication, aligning with the principles of open science.

Furthermore, blockchain technology has the potential to tackle trust issues related to malicious conduct in peer-review procedures [33], deficiencies in study design quality and redundancy [34], as well as limitations on open access to scientific publications. Additionally, the decentralized characteristics of blockchain enable the enhancement of trust in studies and collaborations within intricate scientific projects. Blockchain technology enables either specific groups or the entire network to collectively make decisions regarding projects through regular voting processes, which may adhere to democratic principles, as exemplified by Osgood [35]. The inherent characteristic of immutability (tamper-proof) in blockchain technology perfectly meets the necessity to prevent any form of censorship in open science. In particular, the combination of cryptographic hashing, a consensus mechanism, and decentralization guarantee the immutability of a blockchain. In the context of projects in open science, blockchain can also safeguard intellectual property rights. For instance, it can issue ownership certifications for digital assets stored in a hashed form on the Bitcoin blockchain, as demonstrated in the Bernstein application [36].

Blockchain distinguishes itself from other decentralized systems through its remarkable technical framework, enabling its adaptation to a diverse range of applications. For instance, developers can tailor blockchains to accommodate either open or private access, incorporating individual governance models based on their specific objectives. Beyond the technical aspect, cryptocurrencies offer distinct opportunities, such as the creation of unique business models and incentives for users or entire communities.

Blockchain can enforce immutability and non-repudiation for information stored on it. Regulation and personal humility often stand in the way of this sharing. Blockchain offers new and novel ways to share data securely with only the providers or researchers who are supposed to receive it. For instance, Yates [37] found that the use of the unique identifier generated by the blockchain for private or sensitive data, like medical data, can enhance data security as only the provider and the receiver can see the document. Tlili et al. [38] highlighted the importance of using blockchain to protect data sharing in open science, specifically when generating new Open Educational Resources.

The next section presents the Open REUNICE platform, which harnesses the power of AI and blockchain technologies.

III. Implementation

An innovative open science platform, Open REUNICE (Open research within EUNICE), was developed (http://reunice4u. com/). Open REUNICE harnesses the power of AI and blockchain technologies to foster seamless collaboration among EUNICE partners. Its goal is to establish an environment where researchers can effortlessly connect, share knowledge, and engage in meaningful collaborations. Particularly, 10 universities and 31626 researchers have joined the platform, where 234028 publications and 26 projects were added, accordingly. This platform met the requirements of the EUNICE alliance. Specifically, the project team organized weekly meetings and followed these steps: (1) study of the current situation of open science in the university's alliance; (2) specification of the needs and discussion of the open science problems faced in the alliance; (3) study of feasible and non-feasible functionalities; (4) conception of the project using different diagrams in Unified Modeling Languages; (5) discussion of the design of functionalities between team members; and finally (6) implementation.

To promote a culture of scientific data sharing, this platform was established based on the FAIR principles of open science [39]. Additionally, it seeks to unify the efforts of all universities within the alliance by aligning their open science strategies and implementing a shared platform, fostering collaboration and ensuring mutual benefit.

Based on the analysis of each researcher's profile on the platform, Open REUNICE provides personalized libraries, containing their own publications as well as pertinent research works from their field of expertise. This feature not only streamlines access to relevant information but also encourages interdisciplinary exploration and the discovery of novel research avenues within the alliance. In this context, AI was implemented to provide more accurate resources on the platform. For instance, natural language processing (NLP) was

used to automatically extract keywords from each uploaded resource, analyze them with the research interests of each registered researcher, and finally suggest these resources to those researchers who might be interested in them. Such process helps to build and maintain a research community within Open REUNICE, where researchers could discuss and collaborate on common research topics. Additionally, Blockchain was further used to provide more security and promote the culture of sharing through Open REUNICE. For instance, it gives an identifier for each project, thereby, protecting its intellectual property rights and keeping track of all the updates of contributors. Specifically, the interaction between AI and blockchain on this platform was primarily focused on data management. Blockchain was employed to securely handle and manage the large volumes of data required for training AI models, addressing critical concerns related to data integrity and security [40]. By leveraging blockchain technology, the platform ensures transparent and traceable records of research processes, including AI workflows, data sharing activities, and collaborative efforts. This synergy enhances trust, reproducibility, and accountability in research, aligning with growing calls for more robust frameworks in AI-driven applications [41].

Fig. 1 presents the architecture of Open REUNICE. The system's workflow is color-coded to highlight its key components and processes. Specifically, the red lines represent the process of data collection from various databases and the community-building activities described in detail below. The blue lines depict the analysis process of researchers' publications and projects using AI techniques. Data collected about researchers are stored in the system database, while the outputs of AI-driven analysis—such as extracted keywords and identified research communities—are securely stored on the blockchain to ensure data integrity and security. Finally, the green lines illustrate the safeguarding process for projects. Specifically, all projects are stored on the blockchain system to protect researchers' intellectual property rights and to maintain a transparent record of all contributor updates and changes.
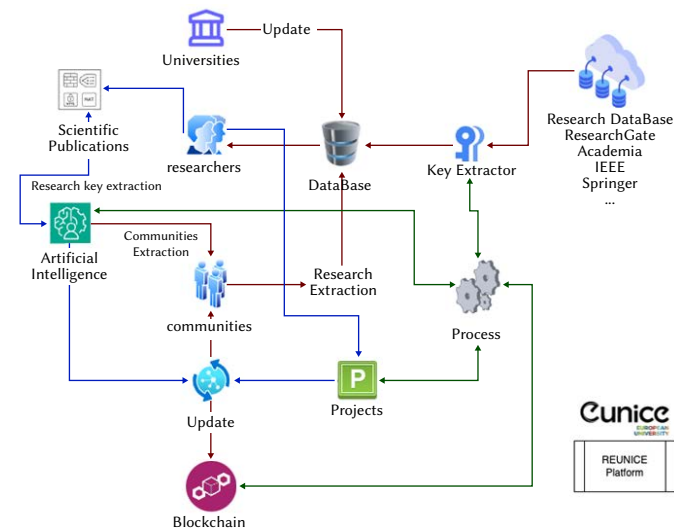
Fig. 1. Open REUNICE platform architecture.

The development process of Open REUNICE was divided into the following four key sequential steps.

A. Data Collection

Open REUNICE uses an application programming interface (API) to extract data about researchers and their publications from ResearchGate, establishing a primary database sourced from research websites. Currently, the platform extracts data from ResearchGate,

with plans to expand its connectivity to other research databases for further information updates in the future (Other APIs can also be added to retrieve more complete representation of researcher's publication history). Specifically, for legal concerns, we consider collecting metadata about publications from ResearchGate and provide links back to the original ResearchGate pages.

Fig. 2 shows the Open REUNICE dashboard of the data collected, illustrating the total number of researchers added to the platform, the contributions made by participants, and the most frequently addressed research topics, from an administrator view point. This comprehensive overview highlights the platform's growth, user engagement, and research focus areas, offering valuable insights into the active involvement of the research community and the trending topics within the platform.

Fig. 2. Screenshot of the dashboard.

B. Building Scientific Community

The researchers' database was constructed based on the collected data from the first step. Specifically, we incorporated information about researchers (university, department, research interests, and publications) into Open REUNICE to create their profiles. Once they join, researchers can update their information and add more publications or projects that they want to share with others. To prevent confusion caused by identical names, researchers' unique identifiers, specifically their open researcher and contributor ID (ORCID), is included in their profiles. The identification of a researcher's ORCID can be done automatically using their email, or manually by the researcher. All collected data about researchers and their research were used by AI to improve collaborations on the platform, as discussed in the next section.

C. Use of AI to Enhance Collaborations and Create Connections

To promote the culture of sharing among researchers from various universities, Open REUNICE relies on AI techniques, specifically NLP, to analyze extensive publication datasets, extracting valuable insights to enhance research quality and discover new patterns and correlations in research trends and methodologies. Given the diverse document types that can be published on the platform, such as research data, scientific articles, and projects, NLP was used to analyze both titles and full texts of researchers' publications, extracting specific research keywords based on publication content (see Fig. 3). The employed technique relies on the frequency of word occurrences in the text. To do so, the YAKE library (Yet Another Keyword Extractor) has been chosen after several tests. YAKE has been recognized in the literature for its efficiency in unsupervised keyword extraction by leveraging statistical text features, such as term frequency, position in the text, and co-occurrence metrics, to identify significant terms without relying on external corpora [42]. The text analysis allows extracting accurate metadata, identifying similarities between publications using these keywords, and categorizing them based on research interests. Specifically, for similarity detection, this platform employs

the Generative Pre-Trained Transformer (GPT-3.5-turbo) model as a pilot test. This model offers a compelling balance of performance and efficiency, aligning with findings in the literature that emphasize the capability of large language models to deliver high accuracy in semantic similarity tasks [43]. Furthermore, GPT-3.5-turbo is designed to be computationally optimized, resulting in lower operational costs and reduced environmental impact compared to earlier iterations of similar models [44].

Furthermore, an AI-based notification system was implemented to inform researchers about new publications and projects on the platform that align with their identified research interests. This system generates automatic recommendations for potential project collaborations. Specifically, if the extracted key words of the new publications and projects match with the keywords of the researcher's publication, a notification will be sent to the researcher. AI was also used to suggest the creation of communities based on matching research interests. This allows creating communities and inviting researchers working on similar topics to join, fostering collaboration and enhancing the overall research environment. This approach not only streamlines the discovery and retrieval of relevant research content but also promotes a more connected and collaborative research community.
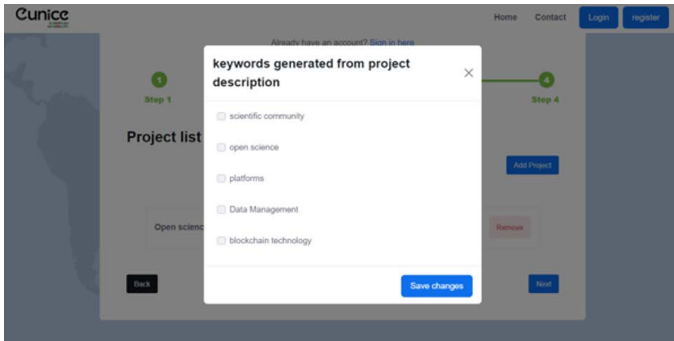
Fig. 3. Screenshots of keyword extraction using NLP.

D. Use of Blockchain to Secure Research Projects Management

To protect intellectual property rights associated with published projects, new ideas, and research data, Open REUNICE was built based on a private Ethereum blockchain. Specifically, it was built using Geth to establish a private network, ensuring the security of all shared data and copyright. In this context, blockchain technology was utilized to issue ownership certificates (in the form of cryptographic hashes) for each project or idea, ensuring secure and verifiable proof of ownership. Fig. 4 presents an example of a project with blockchain reference highlighted in yellow. Blockchain can safeguard researchers' work, ensuring proper attribution and protection against unauthorized use. In addition, each update in the idea by the creator or other contributors is saved in the blockchain to keep track regarding each update (what have been changed). Furthermore, to enhance collaboration, we implemented secure smart contracts on the blockchain to automate and enforce agreements and collaborations about projects among researchers. This reduces administrative overhead and ensures transparent and trustful collaboration. OPEN REUNICE also provides a workspace below the project description where users can leave comments and engage in discussions about the ideas and tasks. Collaborating on projects in a blockchain-based platform allows ensuring the verifiability and persistence of researchers' contributions.
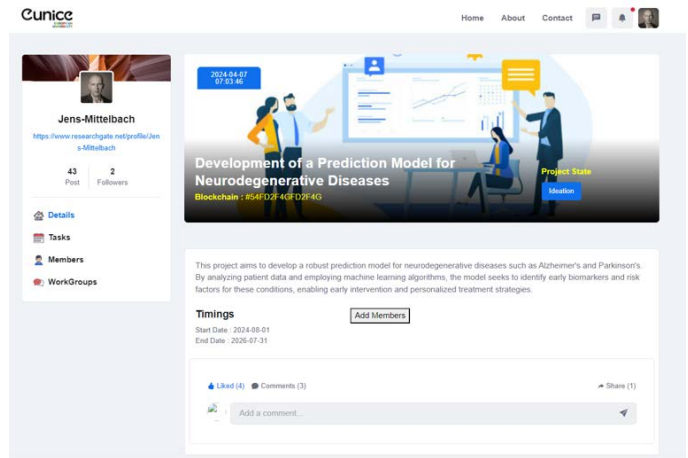
Fig. 4. Screenshot of a project space with blockchain reference highlighted in yellow.

IV. Method

A. Participants and Procedure

Three focus groups with 54 participants were run between 4 June and 25 August 2024 to validate the implemented platform. The first focus group was online with 10 participants, where each participant is a representative from each university of the Alliance. The second focus group with 38 participants, including the 10 participants who assisted in the first focus group, was face to face during the general assembly of the project and it was more open to other researchers from the Alliance. The last focus group was with a sample of 16 active researchers who have used OPEN REUNICE to add their research work and establish various research discussions and collaborations on the platform. Participants were all adults and living or working in European universities. Table I presents the demographic details of participants.

The experiment started with a short presentation of the platform, during which we demonstrated all its functionalities. After that, the first focus group spent around 60 minutes trying and discussing the platform using Zoom, while the second and third focus groups spent around 120 minutes doing the same face-to-face. Even after the controlled experiment concluded, some participants continued to actively use the platform and shared their valuable feedback via email.

TABLE I. Demographic Details of Participants

	Characteristics	N (%)
Gender	Male	18 (33)
	Female	36 (67)
Age	25-35	29 (54)
	35-55	25 (46)
Ethnicity	White-Europe	48 (89)
	White-any other white background	6 (11)
Occupational classification	Researcher	38 (70)
	Administrative occupation	16 (30)

B. Data Collection and Analysis

For data collection, a focus group discussion was conducted, which is a systematic process of collecting data and information on a very specific problem using structured, semi-structured, or unstructured interviews through group discussions [45] [46]. A set of questions was used to facilitate the focus group discussion: (1) "How effectively does the platform support your research or assist you in finding

collaborators working on similar research topics?", (2) "Which features of the platform did you find most useful for your work?", (3) "To what extent do you trust the platform for sharing your innovative ideas and projects?", and (4) "What improvements or additional features would you like to see on the platform?". Specifically, for the online focus group, the session was recorded and transcribed to identify a coding scheme for analyzing focus group interactions, developed by two researchers. After that, collected data was analyzed and coded using this scheme. The coding scheme was further discussed and developed with other researchers during face-to-face research group meetings. The identified codes were refined to answer the aforementioned research questions.

To evaluate the EUNICE community's readiness, we analyzed their attitudes, infrastructure, training, and constraints. The results indicated that EUNICE community has large experience with open science knowledge practices, such as open data and open access, as well as they have all used open science infrastructures, specifically their institutional repositories [39].

V. Results and Discussion

This section presents the results based on the two research questions presented in section I. Specifically, the discussion of the results was based on the benefits of using AI and blockchain in open science platform since no study, to the best of our knowledge, has implemented and discussed the benefits of combining these two technologies in one online platform.

A. RQ1: How Users Perceive the Developed Features in the OPEN EUNICE Platform?

The results showed that overall the participants have positive perceptions toward the implemented features in Open REUNICE. Specifically, the results showed that most of the participants found that all the functionalities are clear, especially researchers. However, some of them highlighted some difficulties in using the platform, specifically in understanding the AI related functionalities. This can be explained by the lack of experience with AI-based open science frameworks, such as semantic scholar, and knowledge in AI technology as many participants are not familiar with computer science and have not received adequate prior training in advance. In this context, several studies stressed the importance of AI literacy for a better use of AI powered technologies [47]. Therefore, it is crucial to provide more training on cutting edge technologies to develop the skills of EUNICE university stakeholders in using Open REUNICE. This strategy aligns with the European Open Science Cloud's (EOSC) comprehensive skills and education strategy, which includes a focus on artificial intelligence [48]. The European Commission's report on digital skills for open science also highlighted the importance of aligning skills with the new Digital Skills Europe objectives. Specifically, the report emphasized that digital skills for FAIR and open science are key enablers for implementing the new European Research Area [48].

Many participants also noted that the use of blockchain technology within the platform was highly beneficial for securely sharing their publications with researchers from other European universities. Additionally, they found it valuable for gaining a comprehensive overview of researchers working on similar topics, their publications, and relevant calls for projects. This enhanced connectivity and transparency not only facilitated collaboration but also fostered a sense of trust and engagement within the research community. Participants emphasized that the platform has the potential to significantly increase the visibility and impact of their publications by sharing them with a wider network of European researchers. Furthermore, it offers opportunities to collaborate and build communities centered around

specific topics of interest. This aligns with the findings of Davis et al. [49], who stated that researchers' perceptions of the value offered by a platform's services play a critical role in shaping their intention to continue using it in the future.

The integration of AI and blockchain features in Open REUNICE encouraged research participants to use it more than one time for sharing their publications. They even indicated an interest in integrating it into their daily activities due to the platform's social features. Users can communicate with others via instant messages, build their own network of friends, view the latest publications from researchers, and use the collaborative workspace to discuss ideas and projects. This result is consistent with the objective of the previous study of Davis et al. [49], which advocate for changing the perception of open science platforms from simple archives to dynamic online homes for materials and collaboration. In this context, the authors claimed that changing the culture of sharing is seen as dependent on changing our view of open science platforms.

B. RQ2: How Can the OPEN REUNICE Platform Address Open Science Challenges, Such as Data Management Issues and Legal Obstacles in the Literature?

One of the primary challenges in open science lies in effective data management and navigating legal issues. This research question explores how Open REUNICE can address these critical challenges, offering potential solutions to enhance the accessibility, security, and compliance of research data within the open science framework.

Results revealed that many participants emphasized the effective use of Open REUNICE in addressing common challenges related to data management, particularly in terms of data findability and reproducibility. For instance, participants appreciated the use of NLP techniques for automatic metadata generation, especially for extracting research keywords directly from publication content. They noted that this approach could produce more accurate metadata, thereby enabling more efficient searches on the platform.

Participants also highlighted that the platform accommodates diverse publication types, including research datasets, scientific articles, and project descriptions, which often have varying structures. This variability can make manual metadata generation challenging. Consequently, many participants valued the platform's AI capabilities for automatically generating metadata based on the content of these documents.

A previous study by Patra et al. [17] also highlighted the challenges of generating automatic metadata for documents with varying structures. To address this issue, they proposed utilizing NLP techniques to generate metadata based on the title and text of the documents. In this context, the literature emphasizes that poor metadata significantly hinders resource retrieval [50]. Furthermore, Leipzig et al. [51] claimed that the generated metadata can affect data reproducibility which can accelerate evaluation and reuse. Another study about Open Educational Resources (OER) found also that the use of AI and NLP techniques can greatly enhance the findability of OER by providing rich and accurate metadata [52].

Additionally, the generated keywords for publications can be generalized and added to a researcher's academic interests, which would aid in creating research communities based on shared interests and provide recommendations based on previous publications. This approach allows researchers to be represented by a set of interests rather than their entire body of work. Many participants believe that such a recommender system could significantly enhance their productivity and expedite further research. In this context, some participants claimed that the proposed research communities for them match with their interests and they were so helpful to discover

the European network of a specific topic. A similar study by Patra et al. [17] developed a recommender system based on automatically generated metadata and found that such a system can significantly improve the reusability of datasets in the biomedical domain. These results align with the general consensus that metadata are crucial in supporting the FAIR principles (Findable, Accessible, Interoperable, Reusable), as demonstrated by the FAIR sharing project [51]. These findings confirm that the integrated AI features in Open REUNICE play a significant role in enhancing data management processes, particularly by improving metadata accuracy, searchability and reusability, thereby addressing critical needs in research workflows.

Regarding the second open science-related challenge that the Open REUNICE can address, the participants expressed that they often face legal challenges when sharing their data, including concerns about copyright infringement and misuse of intellectual property. Blockchain technology can partially solve these issues by providing a secure, transparent, and tamper-proof record of all transactions. For instance, the platform can timestamp and securely store records of intellectual property, ensuring that authorship and contributions are indisputable [53]. Therefore, the participants have indicated that their trust level in sharing their research and data has increased, leading to a more inclusive approach of embracing the culture of sharing. As a result, they expressed their readiness to publish their projects and ideas on the Open REUNICE due to the intellectual property protection offered by blockchain. This is consistent with findings in the literature where blockchain has been shown to significantly enhance trust and transparency in collaborative environments [54]. By securing intellectual property rights, blockchain reduces the fear of idea theft and promotes a more collaborative atmosphere. Researchers are more likely to engage with their peers, share innovative ideas, and build upon each other's work.

Overall, these results confirm that the integration of blockchain technology in Open REUNICE plays a critical role in addressing legal challenges and enhancing user trust. This, in turn, facilitates greater engagement with the platform and supports the broader goals of open science. In addition to the open science challenges identified in the literature, many participants claimed that incorporating a social dimension into Open REUNICE could significantly enhance user engagement. They emphasized that features fostering collaboration, networking, and real-time interaction among researchers could encourage users to interact with the platform on a daily basis, making it a more integral part of their research workflow. Specifically, the entertainment functionalities, such as interactive chats, community forums, and working groups, can make users view engagement with the repository as a normal part of their daily activities. By incorporating these social features, the platform transforms from a simple repository to an interactive community hub, enhancing user experience and fostering a more engaged and collaborative user base. In this context, Sinha et al. [55] found that features like interactive chats and community forums create a sense of community among users, encouraging regular participation. This is particularly important for academic repositories, as increased engagement can lead to more frequent data sharing and collaboration. Additionally, a study by Kimmons and Veletsianos [56] highlighted that social and collaborative features in academic platforms lead to habitual use, as users begin to see the platform as an integral part of their academic workflow. This habitual use not only increases the frequency of data sharing but also enhances the overall user experience.

During the interviews, several participants proposed the integration of emerging technologies that they believed could address challenges identified in the literature or enhance the open science experience. For example, in the context of data reusability—a key aspect of the data management process—participants highlighted the potential of generative AI (GenAI). Given its rapid advancement and increasing utility in academic settings, some researchers suggested leveraging GenAI within Open REUNICE to automatically generate summaries and provide detailed insights into research publications. This functionality could streamline the understanding and dissemination of complex research, thereby promoting greater accessibility and usability of scientific outputs. In this context, a recent study by Hosseini et al. [57], highlighted the potential use of GenAI in open science. For instance, GenAI can simplify complex scientific concepts, eliminate technical jargon, and summarize findings, making research papers easier to understand for non-experts or researchers from different fields. Additionally, GenAI can help improve the identification and connection of diverse outputs, such as data and software, while enhancing their discoverability with more accurate metadata. The results of a recent survey about the use of GenAI by researchers showed that 29% from 3838 researchers are using ChatGPT for finding or summarizing research publications [58].

Furthermore, many participants highlighted the importance of incentives in open science and their potential to motivate researchers to share their research and data on the platform. Some participants emphasized the significance of including researchers' contributions to open science activities in career assessments and personal development processes. This perspective aligns with the European Commission's plans to incorporate open science activities into the research career evaluation system. It is also consistent with the viewpoints of several European universities and organizations, such as the European Open Science Cloud (EOSC), which recognize and value open science activities in research career assessments [48]. In this context, previous studies showed that researchers worried about the impact that open science practices could have on their career since in the traditional evaluation method, researchers are evaluated based on traditional journal metrics and only few incentives from a career perspective to fully commit open science [59]. Allen and Mehler [60] argued that if researchers' contributions to open science, such as preregistrations, the publication of null findings, and the inclusion of DOIs for open code or data are not valued or considered in a scientist's evaluation, then open science will struggle to establish a strong presence in the scientific community. However, different incentive systems should be implemented based on the research field, as some participants suggested, since not all research areas promote open access data. This result aligns with a previous study conducted by Toribio-Flórez et al. [61], which identified that certain open science principles, such as reproducibility, are more prevalent in specific research fields like human and social sciences. In contrast, fields such as artificial intelligence may emphasize promoting open access to data alongside incentive systems that support replication efforts. Since the attitudes of the institutions toward open science will inform the views of its staff, it is very important to align the institutional policies guiding user behavior to change the culture of sharing as highlighted by Davis et al. [49].

VI. Conclusion, Implications, and Future Directions

This study presents a newly developed open science platform Open REUNICE that harnesses the power of AI and blockchain technologies to promote and foster a culture of sharing and seamless collaboration among universities worldwide. The findings from the focus group discussions validated the hypothesis that leveraging AI and blockchain technologies enables researchers and institutions to share open science more efficiently and effectively. Participants highlighted that the platform addresses several key open science challenges frequently noted in the literature, including the automatic generation of metadata, overcoming legal barriers, and addressing social issues. For

instance, the platform's ability to generate automatic metadata based on the publication text can significantly enhance the accuracy and relevance of the metadata, thereby improving content discoverability and retrieval. Additionally, by leveraging blockchain technology, the platform can help mitigate legal barriers related to intellectual property rights, fostering a safer environment for sharing research and innovative ideas. Furthermore, the platform's social features, such as instant messaging, community building, and collaborative workspaces, can promote daily engagement and facilitate a culture of open collaboration among researchers.

In addition to the proposed functionalities of the platform, the participants expressed their desire to add or modify certain features. They suggested incorporating more research databases, such as Scopus and institutional repositories, to gather publications. Additionally, the participants emphasized the importance of implementing an incentive or recognition system for open science activities that can be utilized for career assessments. Using incentives can significantly contribute to creating a robust open science culture by motivating researchers to engage more actively in open science practices.

The results of this study can be used to guide developers in refining and expanding platform features to better meet the needs of the research community, hence promoting the culture of sharing, such as implementing incentive systems. The results can also be used by institutions to develop open science policies to encourage the adoption of open science platforms based on the integration of AI and blockchain technologies to enhance research practices. Furthermore, the platform can serve as a model for international collaborations, demonstrating how technology can overcome common barriers in research sharing and intellectual property protection.

Despite the positive contribution of this study, it still has some limitations that should be acknowledged and further researched. For instance, the number of participants is limited. Additionally, all the participants were from Europe, which can affect sharing various culture perception, since researchers from different places in the world may have different perceptions towards technology based on their background and readiness. Finally, the platform collects data about researchers just from ResarchGate, which can limit the collection of data. Therefore, the future work will focus on overcoming the aforementioned limitations as well as focusing on combining AI and blockchain to overcome the limitations of each technology. For instance, the use of blockchain to overcome AI challenges, such as interpretability issues in AI models.

References

[1] UNESCO, "UNESCO Recommendation on Open Science," *UNESCO*, 2021. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en. [Accessed: 23-Apr-2022].

[2] K. Armeni, L. Brinkman, R. Carlsson, A. Eerland, R. Fijten, R. Fondberg, ... &, R. Zurita-Milla, "Towards wide-scale adoption of open science practices: The role of open science communities," *Science and Public Policy*, vol. 48, no. 5, pp. 605-611, 2021.

[3] R. O. Wright, K. C. Makris, P. Natsiavas, T. Fennell, B. R. Rushing, and A. Wilson, "A long and winding road: culture change on data sharing in exposomics," *Exposome*, vol. 4, no. 1, p. osae004, 2024, doi: 10.1093/exposome/osae004.

[4] S. Leible, S. Schlager, M. Schubotz, and B. Gipp, "A review on blockchain technology and blockchain projects fostering open science," *Frontiers in Blockchain*, vol. 16, 2019, doi: 10.3389/fbloc.2019.00016.

[5] B. Fecher and S. Friesike, "Open Science: One Term, Five Schools of Thought," *in Opening Science*, Cham: Springer, pp. 17-47, 2014, doi: 10.1007/978-3-319-00026-8_2.

[6] C. L. Borgman, "Big Data, Little Data, No Data: Scholarship in the Networked World, Cambridge," MA: MIT Press, 2015, doi: 10.7551/mitpress/9964.001.0001.

[7] C. Tenopir et al., "Changes in data sharing and data reuse practices and perceptions among scientists worldwide," *PLOS ONE*, vol. 10, no. 8, p. e0134826, 2015, doi: 10.1371/journal.pone.0134826.

[8] V. L. Patel and T. G. Kannampallil, "Data Sharing in Healthcare: Ethical and Legal Challenges," *Journal of the American Medical Informatics Association*, vol. 26, no. 4, pp. 403-408, 2019, doi: 10.1093/jamia/ocz038.

[9] K. Wang, "Opportunities in open science with AI," *Frontiers in Big Data*, vol. 2, p. 26, 2019, doi: 10.3389/fdata.2019.00026.

[10] S. Tong, K. Mao, Z. Huang, Y. Zhao, and K. Peng, "Automating Psychological Hypothesis Generation with AI: Large Language Models Meet Causal Graph," *arXiv preprint*, 2024, doi: 10.48550/arXiv.2402.14424.

[11] M. Denden, M. Abed, V. Holotescu, A. Tlili, C. Holotescu, and G. Grosseck, "Down to the rabbit hole: how gamification is integrated in blockchain systems? A systematic literature review," *International Journal of Human–Computer Interaction*, pp. 1–15, 2023.

[12] R. Kumar et al., "Blockchain-based authentication and explainable AI for securing consumer IoT applications," *IEEE Transactions on Consumer Electronics*, 2023, doi: 10.1109/TCE.2023.3287565.

[13] E. Karger, M. Jagals, and F. Ahlemann, "Blockchain for AI data—State of the art and open research," in *Proceedings of the 42nd International Conference on Information Systems (ICIS)*, Austin, TX, USA, pp. 12–15, Dec. 2021.

[14] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016, doi: 10.1038/sdata.2016.18.

[15] L. Zhou, "Patient-centered knowledge sharing in healthcare organizations—Identifying the external barriers," *Informatics for Health & Social Care*, vol. 42, no. 4, pp. 409–420, 2017, doi: 10.1080/17538157.2016.1269106.

[16] A. Zuiderwijk, M. Janssen, and K. Jeffery, "Towards an e-infrastructure to support the provision and use of open data," in *Conference for E-Democracy and Open Government*, pp. 259, May 2013.

[17] B. G. Patra, K. Roberts, and H. Wu, "A content-based dataset recommendation system for researchers—a case study on Gene Expression Omnibus (GEO) repository," *Database*, vol. 2020, p. baaa064, 2020, doi: 10.1093/database/baaa064.

[18] T. Baker and L. Smith, "Educ-AI-tion rebooted? Exploring the future of artificial intelligence in schools and colleges," *Nesta*, 2019. [Online]. Available: https://media.nesta.org.uk/documents/Future_of_AI_and_education_v5_WEB.pdf.

[19] R. Ramachandran, K. Bugbee, and K. Murphy, "From open data to open science," *Earth and Space Science*, vol. 8, no. 5, p. e2020EA001562, 2021, doi: 10.1029/2020EA001562.

[20] O. E. Gundersen, Y. Gil, and D. W. Aha, "On reproducible AI: Towards reproducible research, open science, and digital scholarship in AI publications," *AI Magazine*, vol. 39, no. 3, pp. 56–68, 2018.

[21] E. A. Olivetti et al., "Data-driven materials research enabled by natural language processing and information extraction," *Applied Physics Reviews*, vol. 7, no. 4, 2020.

[22] C. A. Bail, "Can Generative AI Improve Social Science?," Proceedings of the National Academy of Sciences, vol 121, no 21, e2314021121, 2024.

[23] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," *Big Data & Society*, vol. 3, no. 2, p. 2053951716679679, 2016, doi: 10.1177/2053951716679679.

[24] H. F. Atlam, M. A. Azad, A. G. Alzahrani, and G. Wills, "A review of blockchain in Internet of Things and AI," *Big Data and Cognitive Computing*, vol. 4, no. 4, p. 28, 2020, doi: 10.3390/bdcc4040028.

[25] Z. C. Lipton, "The mythos of model interpretability," *arXiv preprint*, 2016. Available: arXiv:1606.03490.

[26] M. Nassar, K. Salah, M. H. ur Rehman, and D. Svetinovic, "Blockchain for explainable and trustworthy artificial intelligence," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 1, p. e1340, 2020, doi: 10.1002/widm.1340.

[27] B. Mons et al., "Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud," *Information Services & Use*, vol. 37, no. 1, pp. 49–56, 2017, doi: 10.3233/ISU-170824.

[28] D. Amodei et al., "Concrete problems in AI safety," *arXiv preprint*, 2016. Available: arXiv:1606.06565.

[29] R. Shinde, S. Patil, K. Kotecha, and K. Ruikar, "Blockchain for securing AI applications and open innovations," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 7, no. 3, p. 189, 2021, doi: 10.3390/joitmc7030189.

[30] A. M. Antonopoulos and G. Wood, *Mastering Ethereum: Building Smart Contracts and DApps*, Sebastopol, CA: O'Reilly Media, 2018.

[31] H. Anwar, "Blockchain vs. Distributed Ledger Technology," 2019. [Online]. Available: https://bit.ly/2SFTRZ0.

[32] C. Furlanello, M. De Domenico, G. Jurman, and N. Bussola, "Towards a scientific blockchain framework for reproducible data analysis," *arXiv preprint*, 2017. Available: arXiv:1707.06552.

[33] M. Dansinger, "Dear plagiarist: a letter to a peer reviewer who stole and published our manuscript as his own," *Annals of Internal Medicine*, vol. 166, no. 2, p. 143, 2017, doi: 10.7326/M16-2414.

[34] J. Belluz and S. Hoffman, "Science is often flawed: it's time we embraced that," 2015.

[35] I. Osgood, "Differentiated products, divided industries: firm preferences over trade liberalization," *Economics & Politics*, vol. 28, no. 2, pp. 161–180, 2016, doi: 10.1111/ecpo.12082.

[36] M. Barulli, F. Weigand, and P. Reboh, "Blockchain solutions for securing intellectual property assets and innovation processes," *Bernstein Product Deck*, 2017. [Online]. Available: https://de.slideshare.net/mbarulli/1702-bernstein-product-deck.

[37] T. D. Yates, "Enhancing healthcare information sharing with blockchain technology," *Open Science Journal*, vol. 5, no. 2, 2020.

[38] A. Tlili et al., "Towards utilising emerging technologies to address the challenges of using Open Educational Resources: a vision of the future," *Educational Technology Research and Development*, vol. 69, pp. 515–532, 2021, doi: 10.1007/s11423-021-09993-3.

[39] M. Denden, "An open science strategy for EUNICE universities," *EUNICE European University*, France, 2023. [Online]. Available: https://eunice-university.eu/research/wp-content/uploads/sites/2/2022/09/REUNICE_DELIVERABLE_3.1.pdf.

[40] P. Zhang, J. White, D. C. Schmidt, and G. Lenz, "Blockchain technology use cases in healthcare and research," *Advances in Computers*, vol. 121, pp. 1–41, 2020, doi: 10.1016/bs.adcom.2020.05.001.

[41] R. Kumar, V. Sharma, and N. Aggarwal, "Blockchain-based frameworks for secure and trustworthy AI systems," *Journal of Emerging Technologies*, vol. 7, no. 2, pp. 89–102, 2021.

[42] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, and A. Jatowt, "YAKE! Keyword extraction from single documents using multiple local features," *Information Sciences*, vol. 509, pp. 257–289, 2020, doi: 10.1016/j.ins.2019.09.013.

[43] T. B. Brown et al., "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165.

[44] D. Patterson et al., "Carbon emissions and large neural network training," *arXiv preprint arXiv:2104.10350*, 2021. [Online]. Available: https://arxiv.org/abs/2104.10350.

[45] Irwanto, *Focused Group Discussion (FGD): Sebuah Pengantar Praktis*, Yayasan Obor Indonesia, 2006. [Online]. Available: https://books.google.co.id/books?id=aRvyDQAAQBAJ&printsec=frontcover&hl=id&source=gbs_atb#v=onepage&q&f=false .

[46] W. Boateng, "Evaluating the efficacy of focus group discussion (FGD) in qualitative social research," *International Journal of Business and Social Science*, vol. 3, no. 7, 2012.

[47] L. Casal-Otero et al., "AI literacy in K-12: a systematic literature review," *International Journal of STEM Education*, vol. 10, no. 1, pp. 29, 2023, doi: 10.1186/s40594-023-00371-x.

[48] European Commission, "Digital skills for FAIR and open science," 2021. [Online]. Available: https://www.ouvrirlascience.fr/wp-content/uploads/2021/02/Digital-Skills-for-FAIR-and-Open-Science.pdf. Accessed on: Jun. 24, 2024

[49] H. C. Davis et al., "Bootstrapping a culture of sharing to facilitate open educational resources," *IEEE Transactions on Learning Technologies*, vol. 3, no. 2, pp. 96–109, 2009, doi: 10.1109/TLT.2009.26.

[50] L. Lannom, D. Koureas, and A. R. Hardisty, "FAIR data and services in biodiversity science and geoscience," *Data Intelligence*, vol. 2, no. 1–2, pp. 122–130, 2020, doi: 10.1162/dint_a_00034.

[51] J. Leipzig, D. Nüst, C. T. Hoyt, K. Ram, and J. Greenberg, "The role of metadata in reproducible computational research," *Patterns*, vol. 2, no. 9, 2021, doi: 10.1016/j.patter.2021.100322.

[52] A. Tlili and D. Burgos, "Unleashing the power of open educational practices (OEP) through artificial intelligence (AI): where to begin?," *Interactive Learning Environments*, pp. 1–8, 2022, doi: 10.1080/10494820.2022.2054603.

[53] D. Tapscott and A. Tapscott, *Blockchain revolution: how the technology behind bitcoin is changing money, business, and the world*, Penguin, 2016.

[54] Y. Zhang, X. Xu, A. Liu, Q. Lu, L. Xu, and F. Tao, "Blockchain-based trust mechanism for IoT-based smart manufacturing system," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 6, pp. 1386–1394, 2019, doi: 10.1109/TCSS.2019.2956470.

[55] A. Sinha, M. M. TK, S. Subramanian, and B. Das, "Text segregation on asynchronous group chat," *Procedia Computer Science*, vol. 171, pp. 1371–1380, 2020, doi: 10.1016/j.procs.2020.04.147.

[56] R. Kimmons and G. Veletsianos, "Education scholars' evolving uses of Twitter as a conference backchannel and social commentary platform," *British Journal of Educational Technology*, vol. 47, no. 3, pp. 445–464, 2016, doi: 10.1111/bjet.12428.

[57] M. Hosseini, S. P. Horbach, K. Holmes, and T. Ross-Hellauer, "Open science at the generative AI turn: An exploratory analysis of challenges and opportunities," 2024.

[58] L. Nordling, "How ChatGPT is transforming the postdoc experience," *Nature*, vol. 622, no. 7983, pp. 655–657, 2023, doi: 10.1038/d41586-023-03235-8.

[59] F. Schönbrodt, "Training students for the open science future," *Nature Human Behaviour*, vol. 3, p. 1031, 2019, doi: 10.1038/s41562-019-0726-z.

[60] C. Allen and D. M. Mehler, "Open science challenges, benefits and tips in early career and beyond," *PLoS Biology*, vol. 17, no. 5, p. e3000246, 2019, doi: 10.1371/journal.pbio.3000246.

[61] D. Toribio-Flórez et al., "Where do early career researchers stand on open science practices? A survey within the Max Planck Society," *Frontiers in Research Metrics and Analytics*, vol. 5, p. 586992, 2021, doi: 10.3389/frma.2020.586992.

Mouna Denden

Dr. Mouna Denden received her Ph.D. degree in computer science from the University of Sfax, Tunisia, in 2020. She is currently a post-doctoral fellow at Université Polytechnique Hauts-de-France (UPHF). She has published several academic papers in refereed international journals and conferences. Ms. Denden has been a member of the local organizing committee and program committees of various international conferences, as well as a reviewer in several peer-reviewed journals. Her current research focuses on educational gamification, educational games, distance learning, learner modeling, adaptive learning, machine learning, human-computer-interactions, educational psychology, open science, artificial intelligence in education and learning analytics.

Mourad Abedn

Prof. Mourad Abed is full Professor in Computer science at the LAMIH UMR CNRS 8201 research lab of the Université Polytechique Hauts-de-France (France). He is Vice-President of Digital technology, Pedagogical Innovation and Strategic and Partnership Projects and the "InnovENT-E Institute" foundation: it aims to support SMEs and SMIs in innovation and internationally. His research activities focus mainly on topics related to knowledge engineering for personalization and information retrieval, design and evaluation of systems in the field of human-machine interaction, social network analysis, semantic modeling and interactive application generation, and design of socially responsible logistic networks. Prof. Abed has graduated over 25 PhD students & HDR and is author and co-author of more than 190 peer reviewed publications in international journals, books, book chapters and conference proceedings. He participates in several research networks, projects, and associations.

Effects of a Flipped Classroom Learning System Integrated With ChatGPT on Students: a Survey From China

Jing Chen[1], Nur Azlina Mohamed Mokmin[1]*, Qi Shen[2]

[1] Centre for Instructional Technology and Multimedia, Universiti Sains Malaysia, Penang (Malaysia)
[2] College of Humanities and Arts, Jiangsu Maritime Institute, Nanjing (China)

* Corresponding author: nurazlina@usm.my

ABSTRACT

In design education, patterns and symbols representing traditional national cultures are often utilized as teaching materials. However, conventional teaching methods frequently fall short in aiding students' comprehension of these intricate symbolisms and abstract concepts, leading to reduced engagement and ineffective learning outcomes. Therefore, we aim to explore whether ChatGPT, as a powerful tool, can assist in solving this problem. Specifically, we integrate ChatGPT into a flipped classroom learning system to assess its effectiveness in enhancing students' understanding of traditional Chinese culture. This research contributes to the feasibility of integrating ChatGPT in design education, particularly in the context of Chinese culture. Additionally, it serves as an exploratory attempt to apply ChatGPT in teaching practices within the field of design.

KEYWORDS

I. INTRODUCTION

ARTIFICIAL Intelligence (AI) continues to advance and optimize, and it plays a crucial role in digital society. AI's ability to process large amounts of data and automate tasks has revolutionized various fields globally [1]. There are increasing indications that AI can have a positive impact on education [2], [3]. However, using generative AI for educational purposes is a relatively new field, and its potential to enhance human learning remains untested mainly [4]. The past two years have seen significant breakthroughs in various generative AI tools, including ChatGPT, developed by OpenAI, which has garnered significant attention worldwide. These advancements have opened up new possibilities for utilizing AI in creative media and educational content, allowing for tasks previously thought to be beyond AI's capabilities [5], [6].

ChatGPT is an advanced conversational AI interface employing Natural Language Processing (NLP) to interact in realistic interactions. This system uses a large-scale language model (LLM) to generate human-like language [7]. It can answer follow-up questions, acknowledge mistakes, challenge incorrect assumptions, and reject inappropriate requests [8]. OpenAI utilizes deep learning algorithms trained on large numbers of texts. These models learn language patterns and structures by processing extensive data and can deliver related and meaningful content to users based on their inquiries [9]. ChatGPT is a versatile tool for various natural language processing tasks, including free-form conversations, text generation, and language translation [10]. It has experienced unprecedented growth, becoming the fastest-growing application in user adoption in history [11].

While ChatGPT presents numerous transformative applications in the field of education, it also introduces a range of challenges and potential threats, leading to mixed reactions among educators [12]. Some educators view AI, such as ChatGPT, as a powerful tool for driving transformative progress in education, while others approach it with skepticism, perceiving it as a possible risk [13]. Farrokhnia [14] performed a SWOT analysis of ChatGPT to evaluate its benefits, including enhanced access to information, personalized learning and reduced instructional workload, as well as its limitations, such as concerns regarding academic integrity, difficulties in assessing response quality, and the potential for bias and discrimination. Adeshola and Adepoju [15] conducted sentiment analysis on ChatGPT, gathering 3,870 usable messages and categorizing them as "positive," "negative," or "neutral." This analysis demonstrated that the majority of participants rated ChatGPT positively. In a review conducted by Pradana [16], existing research on ChatGPT in education was examined through bibliometric analyses and a systematic literature review, affirming its potential for educational applications and highlighting concerns. These debates and concerns around implementing ChatGPT in education underscore the value of conducting thorough analysis and fostering discussions across various realms of education.

As generative artificial intelligence (AI) systems, such as ChatGPT, are increasingly deployed globally, it is essential to integrate principles of Open Science to ensure these tools are accessible, transparent, and inclusive. This requires the open sharing of AI methodologies and data and a concerted effort to adapt AI systems to various regions culturally. Ensuring that AI models are sensitive to diverse cultural contexts is vital for reducing bias and fostering global inclusivity. For instance, Cao et al. [17] observed that while ChatGPT exhibits strong alignment with American cultural norms, its effectiveness diminishes when interacting with users from other cultural backgrounds, highlighting the need for culturally responsive AI systems. Incorporating these considerations into AI development aligns with the goals of Open Science, which advocates for the equitable dissemination of scientific knowledge and technology across different cultural and socio-economic contexts.

Users from diverse cultural backgrounds may encounter linguistic and cultural barriers if the generative AI model does not adequately integrate or learn their culture—a research investigation conducted by Virvou et al. [18] explores the impact of ChatGPT on cultural heritage e-learning. The study highlights the potential of ChatGPT in aiding learners in analyzing and interpreting Greek poetic works. However, it also highlights notable limitations concerning the Greek historical context and factual accuracy. In a feasibility study by Żammit [19] on ChatGPT's assistance in Maltese language learning, the findings reveal limitations in providing information related to Maltese grammar and vocabulary, as well as difficulties in understanding and answering Maltese questions and statements, compared to its effectiveness in aiding the learning of English. The study also notes that 98% of the participants surveyed expressed that ChatGPT lacked cultural context. To our knowledge, no research has explored the feasibility of using ChatGPT to facilitate traditional Chinese cultural learning. Furthermore, there is a scarcity of literature focusing on Chinese students' experiences with ChatGPT, particularly in the context of design education. Addressing this research gap is of utmost importance and should be prioritized.

In this study, we developed a flipped classroom learning system that integrated a ChatGPT-driven pedagogical agent to facilitate student learning in a 2D design course. A quasi-experimental research design was used to assess the impact of this developed system on student learning outcomes, cognitive load, and engagement. To address these objectives, we explored the following research questions.

RQ1. Do Chinese students using a flipped classroom learning system that integrated a ChatGPT-driven pedagogical agent perform better than those using a Basic flipped classroom learning system (i.e., without generative AI components)?

RQ2. Does a flipped classroom learning system that integrates a ChatGPT-driven pedagogical agent affect the cognitive load of Chinese students compared to a Basic flipped classroom learning system?

RQ3. Do Chinese students experience improved engagement when using a flipped classroom learning system that integrated a ChatGPT-driven pedagogical agent, in contrast to a Basic flipped classroom learning system?

II. Background

A. ChatGPT in Education

The performance of ChatGPT varies across different subject areas and is applied in various ways within the field of education [20]. For instance, Kieser et al. [21] conducted research to evaluate the ability of ChatGPT to solve concept inventory (FCI) problems in physics

education. Lee [22] employed ChatGPT as a virtual teaching assistant in medical education, offering students in-depth information and interactive simulations, enhancing student engagement and learning outcomes. Van den Berg and du Plessis [23] investigated the role of ChatGPT in curriculum planning and educational openness within school teacher training. Li [24] utilized ChatGPT as an assistance tool in a courseware project, observing its positive impact on student performance and perception. Moreover, Tlili et al. [25] suggested that future research should incorporate controlled experiments to evaluate the overall effectiveness of ChatGPT in various professional educational settings.

Additionally, students' experiences using ChatGPT for learning may vary due to their different cultural backgrounds and levels of AI literacy. Fui-Hoon et al. [26] highlighted how generative AI, like ChatGPT, can potentially expand the current digital divide in society, raising concerns about equitable access to AI-driven educational resources. They emphasized that the second level of the digital divide, which pertains to the disparity in Internet skills and usage among diverse groups and cultures, has garnered significant attention with the widespread use of the Internet. This disparity challenges open science principles, which advocate for democratizing knowledge and resources. Individuals residing in areas with limited access may face more obstacles in utilizing ChatGPT for assisted learning, particularly if they possess lower AI literacy and less proficient questioning skills [27],[28].

Moreover, as AI tools like ChatGPT continue to evolve, it is crucial to ensure they are culturally adaptive and inclusive to avoid reinforcing biases and to foster a more culturally aware educational environment. Tlili et al. [25] also showed that the results generated by ChatGPT may differ depending on how the questions are asked, even if the dialogue revolves around the same topic. Consequently, learners must employ critical thinking and develop strong questioning skills to achieve optimal outcomes when utilizing ChatGPT, further underscoring the importance of integrating AI literacy into educational practices to bridge these cultural and technological divides.

As mentioned earlier, relevant generative AI in education literature shows the enormous potential of integrating ChatGPT into education to provide timely feedback and assessment, personalize learning experiences, and expand learning resources. However, researchers have also acknowledged several challenges associated with using ChatGPT in education, including potential flaws in cultural context and factual accuracy [18]. Therefore, it is essential to combine ChatGPT with effective instructional design and teaching strategies to facilitate collaborative student learning with the guidance of teachers [16], [25]. Such an approach encourages students to critically analyze and discuss the accuracy of information critically, thereby enhancing their critical thinking skills [13] [18].

B. Flipped Classroom

The flipped classroom is widely recognized as a relevant pedagogical method in educational technology and has been strongly promoted in higher education [29]. In this instructional model, instructors are responsible for providing relevant learning materials such as instructional videos, course websites, and reading texts for students to study before class [30]. Students participate in discussions, group presentations, and additional positive study activities during class time in response to the pre-class materials [31]. After the class period, students are given enriched assignments or quizzes to reinforce their learning [32]. The flipped classroom, as a student-centred learning model, demands learners to be in charge of their study and decision-making throughout the entire process, with the teacher acting as a facilitator [33], [34], [35].

Although positive outcomes have been observed in various forms of flipped classroom development [36], [37], challenges persist [38]. According to some scholars, numerous educators hesitate to embrace the flipped classroom approach because of the additional time and expense required for course adaptation [39]. These include preparing pre-course learning materials, designing learning activities, and managing the classroom environment. Previous research has indicated that student performance within the flipped classroom is primarily influenced by the quality of the pre-class learning materials [40]. Students with a solid grasp of the materials before class are likelier to engage in class and exhibit greater achievement actively. Conversely, students who struggle to comprehend the materials before class may be less engaged in the flipped classroom format. Furthermore, providing personalized instruction to individual students in a flipped learning approach presents a challenging task to teachers [41].

Prior studies on flipped classrooms have focused on their effects on student learning achievement and engagement [32]. Engagement indicates the degree of students' active participation in the learning activity, seeking guidance from the instructor, or collaborating with group members [42], [43]. Behavioural, emotional, cognitive, and agentic engagement are the four student engagement types that promote active classroom learning Behavioral engagement is defined as observable behaviors required for academic achievement; emotional engagement includes how students feel about the learning experience; agentic engagement refers to self-regulated learning conditions; and cognitive engagement refers to applied learning strategies [46]. Several studies have suggested that engagement strongly correlates to academic performance and is a powerful indicator of students' success [47], [48].

Cognitive load theory offers insights into how people adapt to tasks they perform from psychological, physiological, and cognitive perspectives [49]. Cognitive load is the mental effort required to handle and understand information, comprising three main types: intrinsic cognitive load, extrinsic cognitive load, and germane cognitive load [50]. Intrinsic cognitive load relates to the complexity of the learning material, the learner's knowledge base, and their experience level. Extraneous cognitive load relates to the organization and presentation of teaching designs. Effective instructional designs and procedures are essential to mitigate unnecessary cognitive load [51]. Therefore, when designing instruction for the flipped classroom model, it is imperative to consider cognitive load. Although measuring cognitive load is an open question, self-reported mental effort and perceived difficulty are commonly employed as measures of cognitive load in past studies [52], [53].

C. Design Course and Nanjing Yunjin Brocade

The 2D design course involved in this study aims to teach students basic elements, design theories, and various problem-solving techniques. It covers various topics, including design principles, critical thinking, graphic color, and texture. Through this course, students develop their aesthetic and visual concepts and acquire essential skills for visual communication and creative expression [54].

In China, there is a growing emphasis on education in traditional culture, with schools taking on the responsibility of preserving and passing on the country's intangible cultural heritage. To support this, the Chinese government has implemented educational policies promoting heritage education [55]. The university's College of Humanities and Art was recognized by the City of Nanjing in 2019 as an "Intangible Cultural Heritage Education Transmission Base." The college integrates intangible cultural heritage (ICH) into certain curricula, allowing students to relate ICH to their studies and daily lives. Specifically, in the 2D design course, the patterns of Nanjing Intangible Cultural Heritage "Nanjing Yunjin" are examples. This approach enables students to appreciate Nanjing's cultural heritage and unique characteristics while learning graphic design.

In 2009, the craft of Nanjing Yunjin brocade was recognized as being among the "Masterpieces of the Oral and Intangible Heritage of Humanity" [56]. Originating from Nanjing, Nanjing Yunjin brocade derives its name from its delicate patterns resembling clouds in the sky. Its history can be traced back to the third century A.D. it is considered one of China's top three most famous brocades due to its unique and intricate manufacturing techniques. The brocade's vibrant and ever-changing patterns have significant artistic value, drawing elements from traditional Chinese auspicious motifs encompassing animals, plants, and mythological stories. The patterns are created through a fusion of realistic and abstract evolution. Nanjing Yunjin patterns exhibit a finely composed structure, with clear primary and secondary elements and various forms. Designing these patterns requires a deep understanding of 2D design principles, making it a great case study.

However, despite the importance of the 2D design course for art and design students, its conceptual nature and abstract knowledge can be challenging to comprehend. The traditional "chalk and talk" teaching method limits student participation, resulting in low engagement and interest in learning [57]. This often leads to suboptimal teaching outcomes and hampers cultivating students' design foundations. Therefore, we focus on using Nanjing Yunjin brocade as a case study in a 2D design course at a Chinese university. We adopt a flipped classroom model and integrate a ChatGPT-driven teaching agent to support learning. By cultivating aesthetic concepts and understanding basic design principles, we aim to increase students' awareness of intangible cultural heritage and facilitate cultural preservation and development.

III. Flipped Classroom Learning System With ChatGPT

In this study, we designed learning activities using the flipped classroom model and ChatGPT to strengthen students' comprehension of the intangible cultural heritage of the China-Nanjing Yunjin brocade. The structure of our flipped learning system, as illustrated in Fig. 1, consisted of the flipped learning system integrating ChatGPT, the flipped classroom activity management system, and the server database management system. The frontend was constructed using the Vue framework (based on HTML5, CSS, and JavaScript), while the system functionalities were encapsulated as APIs using the Java programming language for frontend interaction. During the implementation process, access to the OpenAI API was obtained, and the corresponding API key was configured in the environment variables of the Linux server. The frontend pages send requests to the backend server, which, upon receiving the requests, invokes the encapsulated APIs to perform Create, Read, Update, and Delete (CRUD) operations on the MySQL database and interact with the OpenAI server. After processing the requests, the server returns the responses to the front end, presenting the processed information to the learners.

Furthermore, the database management system (DBMS) includes learning material, learning profiles, student profiles, and dialogue databases. The learning material database stored learning materials, study sheets, and guidance notes. The student profile database stores students' information, while the learning profile database stores students' learning records. Finally, the dialogue database was utilized to store student dialogue data with ChatGPT.

The content of this study focuses on a two-dimensional design course centered on Nanjing Yunjin brocade. As shown in Fig. 2, students must access the flipped classroom learning system before attending the class to study the learning materials provided. A learning list is available within the system for students to follow. The learning system
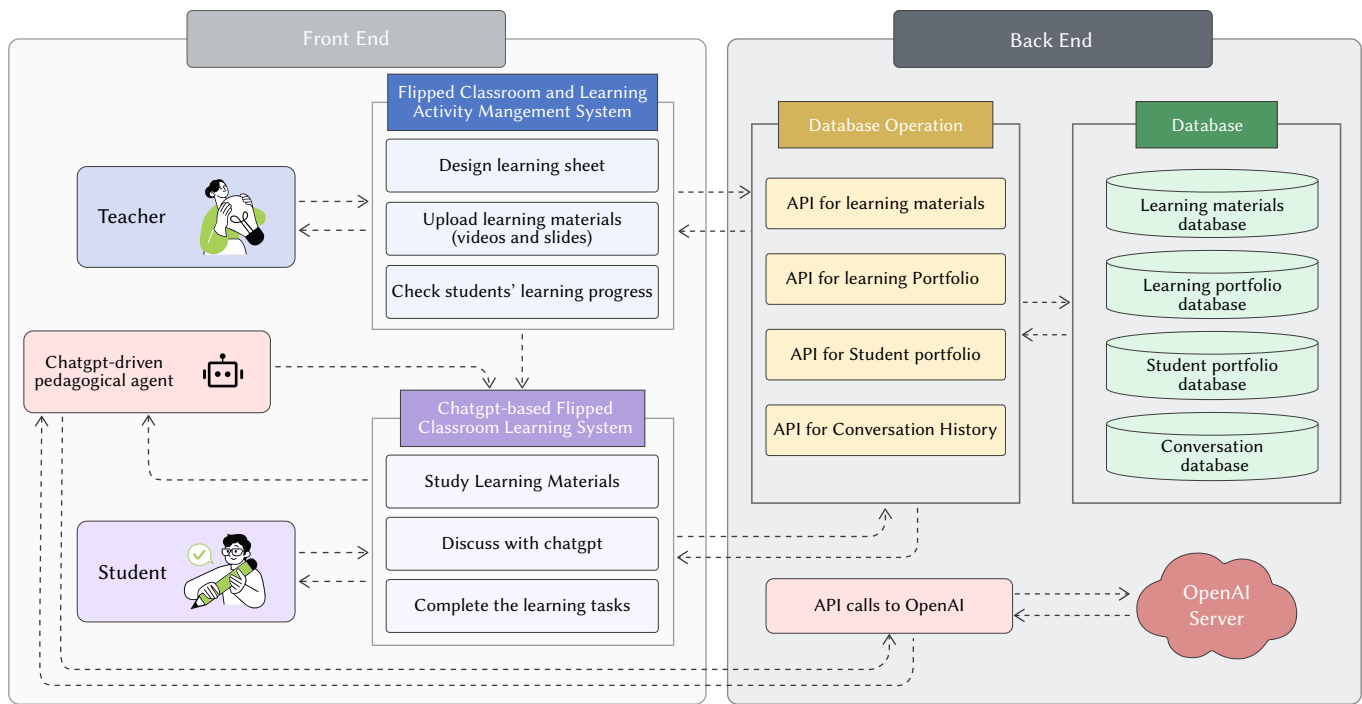
Fig. 1 The structure of Flipped Classroom Learning System with ChatGPT.

Fig. 2 Flipped Classroom Learning System Interface.

offers a range of materials, including slides, learning videos, and text materials. Specifically, slides offer the course's main framework and key points, videos assist in understanding complex concepts through visual presentations, and text materials provide detailed historical background information and relevant academic articles. Students can access and review these materials as often as needed prior to class to gain initial knowledge. Notably, the right side of the interface features a pedagogical agent powered by ChatGPT, designed to assist students in further deepening their understanding and resolving queries. This agent is a standard GPT model whose behavior and generated content are guided by specific prompts. Prompts typically include contextual information or instructions to ensure that the output of the GPT model matches the instructional goals. For example, we have instructed ChatGPT to act as an expert on Nanjing Yunjin brocade in a dialogue, helping students answer questions related to the course content and providing immediate feedback and guidance.

Upon completing the learning content, students must also complete an assignment sheet, which expands on the concepts covered in the learning materials. For instance, they may create a pattern according to Nanjing Yunjin's panel composition principles. Students can discuss with the pedagogical agent powered by ChatGPT at any time and receive personalized references. Furthermore, the flipped classroom learning system emphasizes that ChatGPT is a means of support and that its feedback may not always be correct. Students are encouraged to think critically and can flag any uncertainties to facilitate questions and discussions with classmates and the instructor.

After completing the pre-course learning activities, the teacher organizes group discussions in class, where students can present and discuss their pattern designs. During these discussions, the teacher explains the principles of plate composition and the traditional cultural imagery conveyed. The teacher participates in the discussion and guides the students in response to their doubts, as shown in Fig.3.
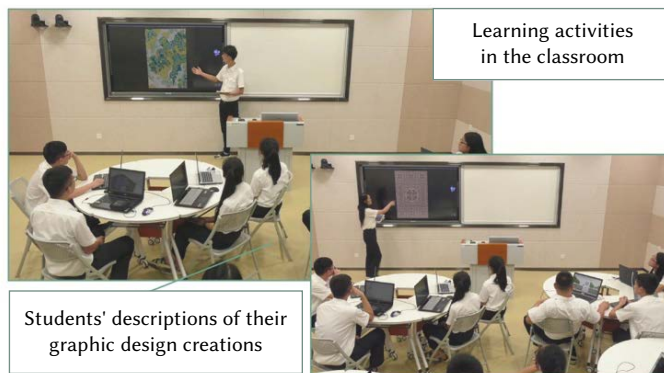
Fig. 3. Learning activities in the classroom.

IV. METHODS

A. Participants

This study involved 70 first-year students from a comprehensive university in China (where the language of instruction was Chinese). Participants were selected from design students enrolled in a 2D design course. Each participant was gifted a Nanjing Yunjin brocade artefact.

In this study, we carefully considered the relevant ethical issues. First, participants were required to read and sign an informed consent form prior to participation. The document outlined the study's purpose, procedures, potential risks and benefits, data usage, and the participants' rights, emphasizing their right to discontinue participation at any time without consequence. Secondly, to safeguard participant privacy, all collected data were anonymized and maintained with strict confidentiality for the exclusive use of this study.

After providing informed consent, participants had to provide personal information, including their name, age, gender, and frequency with which they use ChatGPT (on a 5-point Likert scale). Regarding gender identification, participants were allowed to self-report in fill-in-the-blank questions instead of choosing between male or female options. The mean age of participants was 18.43 years (SD=0.73), with 26 (37.1%) males and 44 (62.9%) females. The participants reported infrequent usage of ChatGPT and similar tools in their daily lives (M=1.84, SD=0.58), and indicated that they mainly never used ChatGPT and similar tools for learning purposes (M=1.34, SD=0.54).

The participants were randomized into experimental (n=32) and control (n=38) groups. This research adopted a quasi-experimental design, incorporating several control variables: (1) Due to the potential influence of teachers' teaching styles, both groups engaged in a flipped classroom approach and were supervised by the same university lecturer with significant experience in design education. (2) All participants were familiarized with the flipped learning system used and informed about the purpose of the research. The experimental group was given additional instruction on how to utilize ChatGPT and associated techniques before the start of the experiment.

B. Measures & Instruments

The instruments used in the study included pre-tests, post-tests, cognitive load items and engagement items. The pre-test and post-test learning scales were designed by three university lecturers with over 12 years of experience teaching two-dimensional design courses and were used to assess student performance. The scale tested students' comprehension of graphic layout and Nanjing Yunjin brocade through 20 multiple-choice questions, each worth 5 points. As students choose the correct answer, their cumulative score ranges from 0 to 100. The scale was assessed by two experts, as shown in Table I.

TABLE I. BACKGROUND AND EXPERIENCE OF THE EXPERTS INVOLVED IN THE EVALUATION

Experts	Background	Experience
A	Inheritor of Nanjing Yunjin brocade weaving skills	41 years of experience in designing Nanjing Yunjin brocade
B	Professor of Art and Design	20 years of experience in art and design research

A questionnaire by Hwang et al. [58] was used to investigate learners' cognitive load during the flipped learning process. This questionnaire includes two dimensions: 'mental load,' which measures learners' intrinsic cognitive load, and 'mental effort,' which assesses learners' extraneous cognitive load. The questionnaire was rated on a six-point Likert scale, of which five items measure mental load and three measure mental effort. In this study, Cronbach's alpha coefficients for the two dimensions were 0.87 and 0.86, respectively.

To evaluate the influence of flipped classroom environment on student engagement, a questionnaire designed by Reeve [44] was utilized. The questionnaire included 21 items and was scored on a five-point Likert scale. The questionnaire measured four dimensions of engagement: "Behavioral Engagement" (5 items) assessed task attention, course engagement, and effort; "Emotional Engagement" (5 items) captured the feelings experienced during learning; "Cognitive Engagement" (4 items) evaluated the development of learning strategies; and "Agentic Engagement" (7 items) appraised students' self-directed learning. For each dimension, Cronbach's alpha coefficients were 0.78, 0.71, 0.73, and 0.79.

The interview questions utilized in this study were adapted from the methodology described by Hwang et al. [59]. They comprised seven questions uniquely crafted to enquire about Chinese students' perceptions of using ChatGPT to facilitate flipped classroom learning activities. All interviews were recorded in audio format to enable comprehensive analysis. Some examples of the interview questions are: "What are your thoughts on this learning system? Could you provide reasons for your opinion?" and "In comparison to your previous experiences with flipped classroom learning, did you notice any differences when using this learning system?"

C. Experimental Procedure

The experimental procedure is depicted in Fig.4. Before the course began, the instructor explained the procedure to all participants to ensure the experiment ran smoothly. Furthermore, all students completed a consent form and a pre-test to assess their foundational understanding of Nanjing Yunjin brocade. Subsequently, both
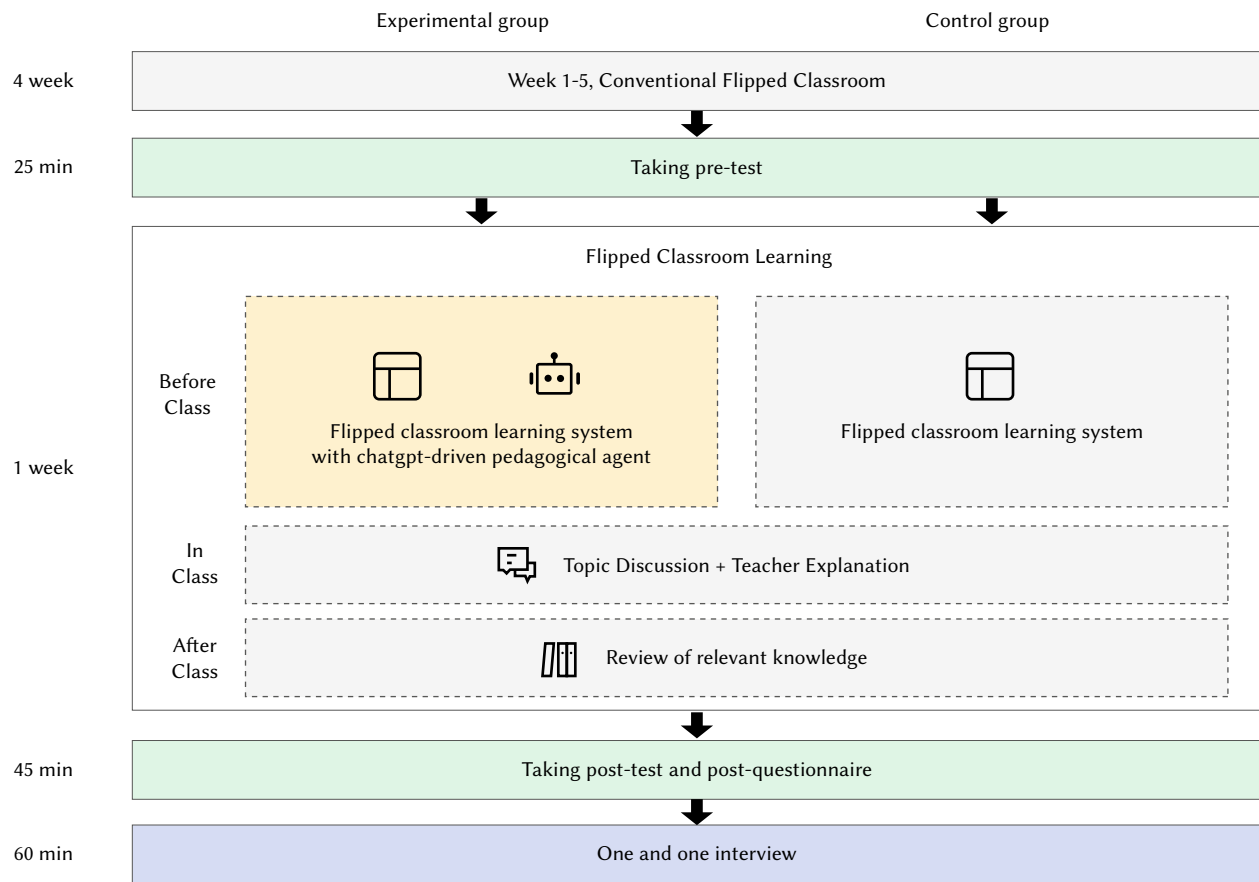
Fig. 4 The procedure of the experiment.

groups utilized the flipped classroom teaching system, with the only difference being the inclusion of ChatGPT. After the learning activity, participants underwent post-tests and completed post-questionnaire regarding engagement and cognitive load. Finally, six randomly selected students from the two groups were interviewed.

V. Results

A. Learning Achievement

A one-way analysis of covariance (ANCOVA) was conducted to examine students' learning achievement. The dependent variable was post-test scores, while the independent variable used two different flipped classroom learning systems for learning activities. Pre-test scores were also included as covariates to account for initial knowledge differences.

A homogeneity test was first performed to validate the ANOVA, which showed that the assumption of regression homogeneity was not violated ($F=2.21$, $p>0.05$). Subsequently, an ANOVA was conducted, and the findings displayed in Table II revealed a significant influence of different flipped classroom learning systems on learning achievement ($F=15.44$, $p<0.001$, $\eta2=0.19$). This indicates that the post-test scores of the two groups varied considerably according to the type of flipped classroom learning system employed. The control group exhibited adjusted mean and standard deviation post-test scores of 70.63 and 10.30, respectively, while the experimental group scored 81.59 and 14.26, respectively. Thus, integrating the ChatGPT-driven pedagogical agent in the flipped classroom learning system improved Chinese students' learning outcomes in 2D design courses, demonstrating a large effect size ($\eta2=0.19$, $\eta2 > 0.14$).

TABLE II. The Analysis of the ANCOVA on Students' Performance

Group	N	Mean	SD	Adjusted mean	F	η2
Experimental group	32	82.5	14.26	81.59	15.44***	0.19
Control group	38	69.87	10.3	70.63		

***$p < .001$

B. Student Engagement

A t-test was conducted to analyze the scores of four engagement subscales (behavioral, agentic, cognitive, and emotional engagement), as presented in Table III, to examine students' engagement. Overall, the experimental group displayed higher mean scores on all subscales, exceeding the median of 3, although with slight variations in each subscale.

Statistically, no significant differences were found between the two groups in behavioral engagement ($t = 1.96$, $p > 0.05$) and agentic engagement ($t = 0.18$, $p > 0.05$). This indicates that including ChatGPT-driven instructional agents did not significantly affect Chinese students' behavioral and agentic engagement in the context of flipped classrooms.

However, for cognitive engagement and emotional engagement, there were significant differences in scores between the two groups: $t = 3.68$ ($p < 0.01$) and $t = 3.07$ ($p < 0.01$), respectively. On the cognitive engagement dimension, the mean values for both groups were 3.68(SD=0.67) and 3.15(SD=0.67), respectively. This suggests that Chinese students using ChatGPT-driven pedagogical agents in the flipped classroom model exhibited higher cognitive engagement than those using the regular flipped classroom learning system. This difference also shows a medium effect size with $d = 0.77$ ($d > 0.50$). Similarly, on the dimension of emotional engagement, the mean values

for both groups were 3.89(SD=0.77) and 3.36(SD=0.66), respectively. This indicates that students in the experimental group demonstrated higher emotional engagement with their studies than the control group. Moreover, this difference shows a medium effect size with d = 0.73 (d > 0.50).

TABLE III. THE T-TEST RESULTS OF STUDENT ENGAGEMENT

	Group	N	Mean	SD	t	d
Behavioral Engagement	Experimental group	32	3.56	0.65	1.96	0.47
	Control group	38	3.28	0.54		
Agentic Engagement	Experimental group	32	3.43	0.72	0.18	0.04
	Control group	38	3.4	0.77		
Cognitive Engagement	Experimental group	32	3.68	0.67	3.21**	0.77
	Control group	38	3.15	0.7		
Emotional Engagement	Experimental group	32	3.89	0.77	3.07**	0.73
	Control group	38	3.36	0.66		

**$p < .01$

C. Cognitive Load

To examine the intrinsic and extraneous cognitive loads experienced by the students, t-tests were conducted to analyze the "mental load" and "mental effort" dimensions. T-tests were conducted to examine the intrinsic and extraneous cognitive loads experienced by the students. As indicated in Table IV, regarding mental load, the t-value was -0.75 ($p > 0.05$) between the two groups. These findings suggest that integrating a ChatGPT-driven pedagogical agent in a flipped classroom learning system does not significantly affect students' mental load compared to the regular flipped classroom learning system. Moreover, no significant difference was found in the level of mental load between the experimental and control groups (t=-1.79, $p < 0.05$). The mean values for the two groups were 3.10 (SD = 0.61) and 3.36 (SD = 0.59), respectively. While the difference is not statistically significant, it is observed that students using the learning system with ChatGPT had slightly lower scores in terms of mental effort compared to the regular learning system. These results imply that incorporating a ChatGPT-driven pedagogical agent in a flipped classroom learning system leads to a slightly reduced extraneous cognitive load during the learning process compared to a regular flipped classroom learning system.

TABLE IV. THE T-TEST RESULT OF THE TWO GROUPS' COGNITIVE LOAD LEVELS

	Group	N	Mean	SD	t	d
Mental Load	Experimental group	32	3.19	0.74	-0.75	0.18
	Control group	38	3.31	0.81		
Mental Effort	Experimental group	32	3.1	0.61	-1.79	0.42
	Control group	38	3.36	0.59		

VI. DISCUSSION

This study aimed to develop a flipped classroom learning system incorporating ChatGPT to enhance a design course for Chinese university students. A quasi-experiment was performed at a university in Nanjing, Jiangsu Province, China, to investigate the effects of using ChatGPT on Chinese students' academic performance,

engagement, and cognitive load. The experiment results demonstrate that combining ChatGPT with a flipped classroom learning system significantly improved students' learning performance and positively influenced their affective and mental engagement. However, the benefits of ChatGPT's integration into educational settings may vary depending on students' cultural backgrounds and levels of AI literacy. Fui-Hoon et al. [26] and others have raised concerns about how AI literacy and digital access disparities may exacerbate educational inequalities. Students from areas with limited access to technology or those with lower AI literacy might struggle to fully leverage the benefits of such advanced tools. This outcome variation underscores the importance of considering these factors in designing and implementing AI-driven educational systems. Notably, it did not significantly affect students' internal or extraneous cognitive loads during the learning activities. The subsequent section will provide a detailed discussion of these findings.

A. Learning Achievement

This study demonstrated that students' academic performance using a flipped classroom learning system incorporating ChatGPT improved, aligning with prior research findings [24], [60]. The system developed in this study aims to provide students with diverse learning materials and immediate personalized guidance. Students can engage in discussions regarding the content with the ChatGPT-driven pedagogical agent and receive tailored feedback. According to Li [24], ChatGPT surpasses other intelligent chatbots in providing learning support, thereby effectively enhancing student performance. Furthermore, students are encouraged to approach the advice supplied by ChatGPT dialectically [20] and discuss them during classroom activities. Through such exchanges, their understanding of the learning content is deepened.

B. Student Engagement

The potency of engagement for learning has been widely researched and is a crucial factor in learning activities, with significant implications for student performance and subject comprehension [61], [62]. In this study, the ChatGPT-driven pedagogical agent positively impacted student engagement. The primary impact was on cognitive, affective, behavioral, and agentic engagement. Cognitive engagement refers to the level of engagement in learning, such as understanding the content [63], [64]. Through interactions with a ChatGPT-driven pedagogical agent, students comprehend and apply what they learn, encouraging critical thinking beyond mere memorization.

Additionally, affective engagement corresponds to emotional responses in education [65], [66]. Flipped classroom learning systems with user-friendly interfaces and ChatGPT-driven pedagogical agents that provide timely feedback and interactions foster a learning environment that promotes student proactivity. Nevertheless, the effectiveness of these interactions can be influenced by the students' cultural contexts and AI literacy. For instance, individuals from different cultural backgrounds might interpret the feedback provided by ChatGPT differently, which could affect their emotional engagement. Moreover, students with varying levels of AI literacy might find it challenging to navigate or fully benefit from these technologies, potentially widening the digital divide within educational settings. However, it needs to be acknowledged that using new technologies in learning may trigger a "novelty effect," which initially enhances engagement and interest [67], [68]. Nonetheless, this effect is transient and diminishes with familiarity with the technology and the associated experience [69], [70], [71].

C. Cognitive Load

Extraneous cognitive load disrupts students' learning and is mainly influenced by how learning materials are presented and how learning

activities are organized [72], [73]. Several previous studies suggest that using chatbots to assist with learning tasks can be a powerful way to decrease extraneous cognitive load [74], [75], [76]. However, students in this study who used ChatGPT-driven pedagogical agents for learning did not experience a significant reduction in extraneous cognitive load when compared to the control group of students. Moreover, based on the information obtained from the interviews, the students reported that there were sometimes significant communication barriers when using the ChatGPT-driven pedagogical agent. Students could not always accurately obtain helpful information when asking ChatGPT questions. As a result, the extraneous cognitive load was not effectively reduced. Three explanations are proposed for this. First, students' proficiency in using ChatGPT was low. ChatGPT generates different results depending on how the questions are asked (e.g., wording), even if the dialogue is about the same topic [77], [25]. Second, ChatGPT is mainly trained with English data, and due to the differences in grammatical structure, expression, and vocabulary between Chinese and English, it may encounter challenges, including inaccuracies or unnatural expressions, when processing Chinese text or generating Chinese content [78]. This is attributed to insufficient Chinese data or understanding Chinese language features. Finally, ChatGPT may not be able to cover all aspects of Chinese culture during training [79], resulting in insufficient comprehension of China-specific cultural, historical, and social contexts.

VII. Conclusion

A. Theoretical and Practical Implications

Our findings have both theoretical and practical implications. Theoretically, we explore the feasibility of using ChatGPT in design education within the Chinese cultural context. While previous research mainly centered on the application of ChatGPT in Western cultural contexts of education [80], [81], [82], this study examines its potential for assisting educational purposes in Chinese cultural contexts, shedding light on the distinct needs and varied experiences of learners from different cultural backgrounds and native languages as they engage with ChatGPT.Our study reveals that, despite being primarily trained on English data, ChatGPT can be effectively utilized in Chinese educational settings with suitable adaptation and localization. This research contributes to the literature on generative AI in multilingual and multicultural contexts by highlighting the necessity of optimizing AI tools for different linguistic environments.

From a practical standpoint, our study incorporates ChatGPT into teaching practice, combining ChatGPT with the flipped classroom teaching model and implementing it within an authentic educational setting. Our study demonstrated that we effectively improve student learning achievement and engagement by integrating ChatGPT into the flipped classroom teaching model. One noteworthy benefit of incorporating ChatGPT in flipped classrooms is that it helps students to ask questions and receive immediate feedback on the content they find challenging to grasp during the pre-class period. This suggests that through rational instructional design, ChatGPT can be a powerful tool to promote teaching traditional Chinese culture in design education, particularly in courses that demand high levels of student interaction and creativity. It offers a reference for exploring the combined application of ChatGPT and teaching strategies. Our research also demonstrated that AI tutors (ChatGPT) collaborate with human teachers to support the achievement of teaching goals. During the pre-classroom learning phase, ChatGPT provides personalized guidance and timely question-and-answer sessions for students. In the classroom, the human teacher facilitates in-depth discussions, Q&A sessions, and practical application of knowledge. This collaborative model informs future educational practices. Additionally, while AI

has been proven to enhance learning experiences across various fields [83]–[85] there is a scarcity of research exploring its feasibility in design education. This study makes a significant contribution by adding new findings to this study area.

B. Limitations and Future Directions

There are several limitations of this research that need to be noted. The first is that the small sample of participants could have influenced the data analysis, highlighting the need for future research with a larger sample size. Secondly, the experiment was conducted within a two-dimensional design course using the Nanjing Yunjin brocade pattern as the teaching case. Therefore, it is important to recognize that different educational areas may yield varying experimental results. Future research should explore the potential impact of ChatGPT within the Chinese cultural context of education.

Although the ChatGPT-driven pedagogical agent demonstrated adequate learning support in this study, the possibility of hallucinations (i.e., generated content containing inaccurate or erroneous information) cannot be overlooked. However, due to the study design limitations, we did not measure the frequency of encountering misinformation or students' reactions. Future research should address this issue by designing experiments to measure and analyze the provision of error messages by ChatGPT and their effects on students' learning behaviors and outcomes.

Furthermore, variations in participants' familiarity with AI tools in this study may have influenced their learning performance and experimental results. Future research should measure and control participants' familiarity with AI tools, exploring how to achieve equitable learning outcomes among students with varying levels of familiarity.

Lastly, it is imperative to emphasize that using a ChatGPT-driven pedagogical agent in this study inevitably poses the risk of encountering harmful behaviors such as dishonesty, manipulation, and misinformation. Future research should focus on designing and implementing safer and more suitable chatbots as pedagogical agents and, for instance, integrating a real-time user feedback system so that students and instructors can report errors or inappropriate information promptly and optimize the performance of the pedagogical agent through the collected feedback.

References

[1] W. Yang, "Artificial Intelligence education for young children: Why, what, and how in curriculum design and implementation," *Computers and Education: Artificial Intelligence*, vol. 3, pp. 100061, 2022, doi: 10.1016/J.CAEAI.2022.100061.

[2] A. Y. Q. Huang, O. H. T. Lu, and S. J. H. Yang, "Effects of artificial Intelligence–Enabled personalized recommendations on learners' learning engagement, motivation, and outcomes in a flipped classroom," *Computers and Education*, vol. 194, 2023, doi: 10.1016/J.COMPEDU.2022.104684.

[3] C. Vallis, S. Wilson, D. Gozman, and J. Buchanan, "Student Perceptions of AI-Generated Avatars in Teaching Business Ethics: We Might not be Impressed," *Postdigital Science and Education*, pp. 1–19, 2023, doi: 10.1007/S42438-023-00407-7.

[4] D. Leiker, A. R. Gyllen, I. Eldesouky, and M. Cukurova, "Generative AI for learning: Investigating the potential of synthetic learning videos," 2023. Available: https://arxiv.org/abs/2304.03784v2.

[5] D. Baidoo-Anu and L. Owusu Ansah, "Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning," *SSRN Electronic Journal*, 2023, doi: 10.2139/SSRN.4337484.

[6] Z. Epstein, A. Hertzmann, M. Akten, H. Farid, J. Fjeld, M. R. Frank, M. Groh, L. Herman, N. Leach, R. Mahari, A. Pentland, O. Russakovsky, H. Schroeder, and A. Smith, "Art and the science of generative AI," *Science*

(New York, N.Y.), vol. 380, no. 6650, pp. 1110–1111, 2023, doi: 10.1126/SCIENCE.ADH4451.

[7] OpenAI, "OpenAI," 2023. Available: https://openai.com/.

[8] OpenAI, "Introducing ChatGPT," 2023. Available: https://openai.com/blog/chatgpt.

[9] D. Zhou, S. Liu, and S. Grassini, "Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings," *Education Sciences*, vol. 13, no. 7, pp. 692, 2023, doi: 10.3390/EDUCSCI13070692.

[10] M. M. Rahman and Y. Watanobe, "ChatGPT for Education and Research: Opportunities, Threats, and Strategies," *Applied Sciences*, vol. 13, no. 9, pp. 5783, 2023, doi: 10.3390/APP13095783.

[11] H. Krystal, "ChatGPT sets record for fastest-growing user base - analyst note | Reuters," 2023, February 2. Available: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[12] M. Montenegro-Rueda, J. Fernández-Cerero, J. M. Fernández-Batanero, & E. López-Meneses, "Impact of the Implementation of ChatGPT in Education: A Systematic Review," *Computers*, vol. 12, no. 8, p. 153, 2023, doi: 10.3390/COMPUTERS12080153.

[13] R. Hadi Mogavi, C. Deng, J. Juho Kim, P. Zhou, Y. D. Kwon, A. Hosny Saleh Metwally, A. Tlili, S. Bassanelli, A. Bucchiarone, S. Gujar, L. E. Nacke, & P. Hui, "ChatGPT in education: A blessing or a curse? A qualitative study exploring early adopters' utilization and perceptions," *Computers in Human Behavior: Artificial Humans*, vol. 2, no. 1, p. 100027, 2024, doi: 10.1016/J.CHBAH.2023.100027.

[14] M. Farrokhnia, S. K. Banihashem, O. Noroozi, & A. Wals, "A SWOT analysis of ChatGPT: Implications for educational practice and research," *Innovations in Education and Teaching International*, pp. 1–15, 2023, doi: 10.1080/14703297.2023.2195846.

[15] I. Adeshola & A. P. Adepoju, "The opportunities and challenges of ChatGPT in education," *Interactive Learning Environments*, pp. 1–14, 2023, doi: 10.1080/10494820.2023.2253858.

[16] M. Pradana, H. P. Elisa, & S. Syarifuddin, "Discussing ChatGPT in education: A literature review and bibliometric analysis," *Cogent Education*, vol. 10, no. 2, 2023, doi: 10.1080/2331186X.2023.2243134.

[17] Y. Cao, L. Zhou, S. Lee, L. Cabello, M. Chen, & D. Hershcovich, "Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study," in *EACL 2023 - Cross-Cultural Considerations in NLP @ EACL, Proceedings of the Workshop*, pp. 53–67, 2023, doi: 10.18653/v1/2023.c3nlp-1.7.

[18] M. Virvou, G. A. Tsihrintzis, D. N. Sotiropouloss, K. Chrysafiadi, E. Sakkopoulos, & E.-A. Tsichrintzi, "ChatGPT in Artificial Intelligence-Empowered E-Learning for Cultural Heritage: The case of Lyrics and Poems," pp. 1–9, 2023, doi: 10.1109/IISA59645.2023.10345878.

[19] J. Żammit, "Harnessing the Power of ChatGPT for Mastering the Maltese Language: A Journey of Breaking Barriers and Charting New Paths," *Studies in Computational Intelligence*, vol. 1105, pp. 161–178, 2023, doi: 10.1007/978-3-031-37454-8_8/COVER.

[20] C. K. Lo, "What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature," *Education Sciences*, vol. 13, no. 4, p. 410, 2023, doi: 10.3390/educsci13040410

[21] F. Kieser, P. Wulff, J. Kuhn, & S. Küchemann, "Educational data augmentation in physics education research using ChatGPT," *Physical Review Physics Education Research*, vol. 19, no. 2, p. 020150, 2023, doi: 10.1103/PhysRevPhysEducRes.19.020150.

[22] H. Lee, "The rise of ChatGPT: Exploring its potential in medical education," *Anatomical Sciences Education*, 2023, doi: 10.1002/ASE.2270.

[23] G. van den Berg & E. du Plessis, "ChatGPT and Generative AI: Possibilities for Its Contribution to Lesson Planning, Critical Thinking and Openness in Teacher Education," *Education Sciences*, vol. 13, no. 10, p. 998, 2023, doi: 10.3390/EDUCSCI13100998.

[24] H. Li, "Effects of a ChatGPT-based flipped learning guiding approach on learners' courseware project performances and perceptions," *Australasian Journal of Educational Technology*, vol. 39, no. 5, pp. 40–58, 2023, doi: 10.14742/ajet.8923.

[25] A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, & B. Agyemang, "What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education," *Smart Learning Environments*, vol. 10, no. 1, pp. 1–24, 2023, doi: 10.1186/S40561-023-00237-X.

[26] F. H. Nah, R. Zheng, J. Cai, K. Siau, & L. Chen, "Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration," *Journal of Information Technology Case and Application Research*, vol. 25, no. 3, pp. 277–304, 2023, doi: 10.1080/15228053.2023.2233814.

[27] A. Bozkurt & R. C. Sharma, "Challenging the Status Quo and Exploring the New Boundaries in the Age of Algorithms: Reimagining the Role of Generative AI in Distance Education and Online Learning," *Asian Journal of Distance Education*, vol. 18, no. 1, 2023. Available: https://asianjde.com/ojs/index.php/AsianJDE/article/view/714.

[28] L. Carter, D. Liu, & C. Cantrell, "Exploring the Intersection of the Digital Divide and Artificial Intelligence: A Hermeneutic Literature Review," *AIS Transactions on Human-Computer Interaction*, vol. 12, no. 4, pp. 253–275, 2020, doi: 10.17705/1thci.00138.

[29] P. Strelan, A. Osborn, & E. Palmer, "The flipped classroom: A meta-analysis of effects on student performance across disciplines and education levels," *Educational Research Review*, vol. 30, p. 100314, 2020, doi: 10.1016/J.EDUREV.2020.100314.

[30] G. Akçayır & M. Akçayır, "The flipped classroom: A review of its advantages and challenges," *Computers & Education*, vol. 126, pp. 334–345, 2018, doi: 10.1016/J.COMPEDU.2018.07.021.

[31] Meyliana, B. Sablan, S. Surjandy, & A. N. Hidayanto, "Flipped learning effect on classroom engagement and outcomes in university information systems class," *Education and Information Technologies*, vol. 27, no. 3, pp. 3341–3359, 2022, doi: 10.1007/s10639-021-10723-9.

[32] H. Al-Samarraie, A. Shamsuddin, & A. I. Alzahrani, "A flipped classroom model in higher education: a review of the evidence across disciplines," *Educational Technology Research and Development*, vol. 68, no. 3, pp. 1017–1051, 2020, doi: 10.1007/s11423-019-09718-8.

[33] R. Brewer & S. Movahedazarhouligh, "Successful stories and conflicts: A literature review on the effectiveness of flipped learning in higher education," *Journal of Computer Assisted Learning*, vol. 34, no. 4, pp. 409–416, 2018, doi: 10.1111/JCAL.12250.

[34] C. Jia, K. F. Hew, J. Diahui, & L. Liuyufeng, "Towards a fully online flipped classroom model to support student learning outcomes and engagement: A 2-year design-based study," *The Internet and Higher Education*, vol. 56, p. 100878, 2023, doi: 10.1016/J.IHEDUC.2022.100878.

[35] M. K. Kim, S. M. Kim, O. Khera, & J. Getman, "The experience of three flipped classrooms in an urban university: an exploration of design principles," *The Internet and Higher Education*, vol. 22, pp. 37–50, 2014, doi: 10.1016/J.IHEDUC.2014.04.003.

[36] L. Cheng, A. D. Ritzhaupt, & P. Antonenko, "Effects of the flipped classroom instructional strategy on students' learning outcomes: a meta-analysis," *Educational Technology Research and Development*, vol. 67, no. 4, pp. 793–824, 2019, doi: 10.1007/s11423-018-9633-7.

[37] J. Lee, C. Lim, & H. Kim, "Development of an instructional design model for flipped learning in higher education," *Educational Technology Research and Development*, vol. 65, no. 2, pp. 427–453, 2017, doi: 10.1007/S11423-016-9502-1.

[38] K. Missildine, R. Fountain, L. Summers, & K. Gosselin, "Flipping the classroom to improve student performance and satisfaction," *Journal of Nursing Education*, vol. 52, no. 10, pp. 597–599, 2013, doi: 10.3928/01484834-20130919-03.

[39] T. Wang, "Overcoming barriers to 'flip': building teacher's capacity for the adoption of flipped classroom in Hong Kong secondary schools," *Research and Practice in Technology Enhanced Learning*, vol. 12, no. 1, p. 6, 2017, doi: 10.1186/s41039-017-0047-7.

[40] M. Förster, A. Maur, C. Weiser, & K. Winkel, "Pre-class video watching fosters achievement and knowledge retention in a flipped classroom," *Computers & Education*, vol. 179, p. 104399, 2022, doi: 10.1016/J.COMPEDU.2021.104399.

[41] S. C. Chang & G. J. Hwang, "Impacts of an augmented reality-based flipped learning guiding approach on students' scientific project performance and perceptions," *Computers & Education*, vol. 125, pp. 226–239, 2018, doi: 10.1016/J.COMPEDU.2018.06.007.

[42] M. Bond, "Facilitating student engagement through the flipped learning approach in K-12: A systematic review," *Computers & Education*, vol. 151, p. 103819, 2020, doi: 10.1016/J.COMPEDU.2020.103819.

[43] M. Bond, K. Buntins, S. Bedenlier, O. Zawacki-Richter, & M. Kerres, "Mapping research in student engagement and educational technology in higher education: a systematic evidence map," *International Journal of*

Educational Technology in Higher Education, vol. 17, no. 1, p. 1, 2020, doi: 10.1186/S41239-019-0176-8.

[44] J. Reeve, "How students create motivationally supportive learning environments for themselves: The concept of agentic engagement," *Journal of Educational Psychology*, vol. 105, no. 3, pp. 579–595, 2013, doi: 10.1037/A0032690.

[45] J. Reeve, H. Jang, D. Carrell, S. Jeon, & J. Barch, "Enhancing students' engagement by increasing teachers' autonomy support," *Motivation and Emotion*, vol. 28, no. 2, pp. 147–169, 2004, doi: 10.1023/B:MOEM.0000032312.95499.6F.

[46] J. Reeve, S. H. Cheon, & H. Jang, "How and why students make academic progress: Reconceptualizing the student engagement construct to increase its explanatory power," *Contemporary Educational Psychology*, vol. 62, p. 101899, 2020, doi: 10.1016/J.CEDPSYCH.2020.101899.

[47] A. Y. Q. Huang, O. H. T. Lu, & S. J. H. Yang, "Effects of artificial Intelligence–Enabled personalized recommendations on learners' learning engagement, motivation, and outcomes in a flipped classroom," *Computers & Education*, vol. 194, p. 104684, 2023, doi: 10.1016/j.compedu.2022.104684.

[48] E. A. Skinner & J. R. Pitzer, "Developmental dynamics of student engagement, coping, and everyday resilience," in *Handbook of Research on Student Engagement*, pp. 21–44, 2012, doi: 10.1007/978-1-4614-2018-7_2/COVER.

[49] W. Schnotz & C. Kürschner, "A reconsideration of cognitive load theory," *Educational Psychology Review*, vol. 19, no. 4, pp. 469–508, 2007, doi: 10.1007/S10648-007-9053-4.

[50] M. Klepsch & T. Seufert, "Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load," *Instructional Science*, vol. 48, no. 1, pp. 45–77, 2020, doi: 10.1007/S11251-020-09502-9/TABLES/11.

[51] J. L. Plass & S. Kalyuga, "Four Ways of Considering Emotion in Cognitive Load Theory," *Educational Psychology Review*, vol. 31, no. 2, pp. 339–359, 2019, doi: 10.1007/S10648-019-09473-5/METRICS.

[52] T. de Jong, "Cognitive load theory, educational research, and instructional design: Some food for thought," *Instructional Science*, vol. 38, no. 2, pp. 105–134, 2010, doi: 10.1007/S11251-009-9110-0/METRICS.

[53] J. Sweller, "The Role of Independent Measures of Load in Cognitive Load Theory," in *Cognitive Load Measurement and Application*, pp. 3–7, 2017, doi: 10.4324/9781315296258-1.

[54] N. A. M. Mokmin, Hanjun, S., Jing, C., & Qi, S., "Impact of an AR-based learning approach on the learning achievement, motivation, and cognitive load of students on a design course," *Journal of Computers in Education*, 2023, doi: 10.1007/s40692-023-00270-2.

[55] A. P. Underhill & L. C. Salazar, Eds., *Finding Solutions for Protecting and Sharing Archaeological Heritage Resources*, 2016, doi: 10.1007/978-3-319-20255-6.

[56] UNESCO, "Craftsmanship of Nanjing Yunjin brocade," 2009. Available: https://ich.unesco.org/en/RL/craftsmanship-of-nanjing-yunjin-brocade-00200.

[57] N. Singh, "'A Little Flip Goes a Long Way'—The Impact of a Flipped Classroom Design on Student Performance and Engagement in a First-Year Undergraduate Economics Classroom," *Education Sciences*, vol. 10, no. 11, p. 319, 2020, doi: 10.3390/EDUCSCI10110319.

[58] G. J. Hwang, L. H. Yang, & S. Y. Wang, "A concept map-embedded educational computer game for improving students' learning performance in natural science courses," *Computers & Education*, vol. 69, pp. 121–130, 2013, doi: 10.1016/J.COMPEDU.2013.07.008.

[59] G. J. Hwang, T. C. Yang, C. C. Tsai, & S. J. H. Yang, "A context-aware ubiquitous learning environment for conducting complex science experiments," *Computers & Education*, vol. 53, no. 2, pp. 402–413, 2009, doi: 10.1016/J.COMPEDU.2009.02.016.

[60] C.-J. Lin & H. Mubarok, "Learning Analytics for Investigating the Mind Map-Guided AI Chatbot Approach in an EFL Flipped Speaking Classroom," *Educational Technology & Society*, vol. 24, no. 4, pp. 16–35, 2021, Available: https://www.jstor.org/stable/48629242.

[61] A. Martinez-Lincoln, M. A. Barnes, & N. H. Clemens, "The influence of student engagement on the effects of an inferential reading comprehension intervention for struggling middle school readers," *Annals of Dyslexia*, vol. 71, no. 2, pp. 322–345, 2021, doi: 10.1007/s11881-020-00209-7.

[62] Y. B. Rajabalee & M. I. Santally, "Learner satisfaction, engagement and performances in an online module: Implications for institutional e-learning policy," *Education and Information Technologies*, vol. 26, no. 3, pp. 2623–2656, 2021, doi: 10.1007/S10639-020-10375-1/FIGURES/5.

[63] I. Buil, S. Catalán, & E. Martínez, "Engagement in business simulation games: A self-system model of motivational development," *British Journal of Educational Technology*, vol. 51, no. 1, pp. 297–311, 2020, doi: 10.1111/BJET.12762.

[64] L. Ding, C. M. Kim, & M. Orey, "Studies of student engagement in gamified online discussions," *Computers & Education*, vol. 115, pp. 126–142, 2017, doi: 10.1016/J.COMPEDU.2017.06.016.

[65] J. A. Fredricks, P. C. Blumenfeld, & A. H. Paris, "School Engagement: Potential of the Concept, State of the Evidence," *Review of Educational Research*, vol. 74, no. 1, pp. 59–109, 2004, doi: 10.3102/00346543074001059.

[66] J. Reeve & C. M. Tseng, "Agency as a fourth aspect of students' engagement during learning activities," *Contemporary Educational Psychology*, vol. 36, no. 4, pp. 257–267, 2011, doi: 10.1016/J.CEDPSYCH.2011.05.002.

[67] S. G. Fussell, J. L. Derby, J. K. Smith, W. J. Shelstad, J. D. Benedict, B. S. Chaparro, R. Thomas, & A. R. Dattel, "Usability Testing of a Virtual Reality Tutorial," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63, no. 1, pp. 2303–2307, 2019, doi: 10.1177/1071181319631494.

[68] I. Miguel-Alonso, B. Rodriguez-Garcia, D. Checa, & A. Bustillo, "Countering the Novelty Effect: A Tutorial for Immersive Virtual Reality Learning Environments," *Applied Sciences*, vol. 13, no. 1, p. 593, 2023, doi: 10.3390/app13010593.

[69] M. Koch, K. von Luck, J. Schwarzer, & S. Draheim, "The Novelty Effect in Large Display Deployments – Experiences and Lessons-Learned for Evaluating Prototypes," 2018. Available: https://doi.org/10.18420/ECSCW2018_3.

[70] L. Rodrigues, F. D. Pereira, A. M. Toda, P. T. Palomino, M. Pessoa, L. S. G. Carvalho, D. Fernandes, E. H. T. Oliveira, A. I. Cristea, & S. Isotani, "Gamification suffers from the novelty effect but benefits from the familiarization effect: Findings from a longitudinal study," *International Journal of Educational Technology in Higher Education*, vol. 19, no. 1, p. 13, 2022, doi: 10.1186/s41239-021-00314-6.

[71] C. H. Tsay, A. K. Kofinas, S. K. Trivedi, & Y. Yang, "Overcoming the novelty effect in online gamified learning systems: An empirical evaluation of student engagement and performance," *Journal of Computer Assisted Learning*, vol. 36, no. 2, pp. 128–146, 2020, doi: 10.1111/jcal.12385.

[72] W. Atiomo, "Cognitive load theory and differential attainment," *BMJ*, vol. 368, 2020, doi: 10.1136/BMJ.M965.

[73] W. Leahy & J. Sweller, "Cognitive load theory and the effects of transient information on the modality effect," *Instructional Science*, vol. 44, no. 1, pp. 107–123, 2016, doi: 10.1007/S11251-015-9362-9/METRICS.

[74] T. Li, Y. Ji, & Z. Zhan, "Expert or machine? Comparing the effect of pairing student teacher with in-service teacher and ChatGPT on their critical thinking, learning performance, and cognitive load in an integrated-STEM course," *Asia Pacific Journal of Education*, 2024, doi: 10.1080/02188791.2024.2305163.

[75] J. Schmidhuber, S. Schlogl, & C. Ploder, "Cognitive Load and Productivity Implications in Human-Chatbot Interaction," *Proceedings of the 2021 IEEE International Conference on Human-Machine Systems, ICHMS 2021*, 2021, doi: 10.1109/ICHMS53169.2021.9582445.

[76] T. T. Wu, H. Y. Lee, P. H. Li, C. N. Huang, & Y. M. Huang, "Promoting Self-Regulation Progress and Knowledge Construction in Blended Learning via ChatGPT-Based Learning Aid," *Journal of Educational Computing Research*, vol. 61, no. 8, pp. 3–31, 2023, doi: 10.1177/07356331231191125.

[77] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, & V. Tsengid, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digital Health*, vol. 2, no. 2, e0000198, 2023, doi: 10.1371/JOURNAL.PDIG.0000198.

[78] L. Li, H. Zhang, C. Li, H. You, & W. Cui, "Evaluation on ChatGPT for Chinese Language Understanding," *Data Intelligence*, vol. 5, no. 4, pp. 885–903, 2023, doi: 10.1162/dint_a_00232.

[79] M. Liu, Y. Ren, L. M. Nyagoga, F. Stonier, Z. Wu, & L. Yu, "Future of education in the era of generative artificial intelligence: Consensus among Chinese scholars on applications of ChatGPT in schools," *Future in Educational Research*, vol. 1, no. 1, pp. 72–101, 2023, doi: 10.1002/fer3.10.

[80] G. Cooper, "Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence," *Journal of Science Education and Technology*, vol. 32, no. 3, pp. 444–452, 2023, doi: 10.1007/S10956-023-10039-Y/TABLES/1.

[81] J.-N. García-Sánchez, H.-C. K. Lin, J. S. Jauhiainen, & A. G. Guerra, "Generative AI and ChatGPT in School Children's Education: Evidence from a School Lesson," *Sustainability*, vol. 15, no. 18, p. 14025, 2023, doi: 10.3390/SU151814025.

[82] A. Gilson, C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, & D. Chartash, "How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment," *JMIR Medical Education*, vol. 9, no. 1, e45312, 2023, doi: 10.2196/45312.

[83] N. A. M. Mokmin & M. Masood, "Case-based reasoning and profiling system for learning mathematics (CBR-PROMATH)," in *Lecture Notes in Electrical Engineering* (Vol. 315, pp. 939–948), 2015, doi: 10.1007/978-3-319-07674-4_88.

[84] N. A. M. Mokmin & M. Masood, "The design and development of an intelligent tutoring system as a part of the architecture of the Internet of Things (IoT)," in *ACM International Conference Proceeding Series* (Vol. 2017-Octob), 2017, doi: 10.1145/3145777.3145793.

[85] S. Shahriar & K. Hayawi, "Let's Have a Chat! A Conversation with ChatGPT: Technology, Applications, and Limitations," *Artificial Intelligence Applications*, pp. 1–16, 2023, doi: 10.47852/bonviewaia3202939.

Jing Chen

She is a PhD student at the Centre for Instructional Technology and Multimedia, University of Science Malaysia (USM). She obtained her degree in Fine Arts from Jinling Institute of Science and Technology (JIST) and her Master of Fine Arts (MFA) from Soochow University of Science and Technology (SUST). During her master's degree, she was involved in research projects related to virtual reality applications and intangible cultural heritage, and her current doctoral research involves art and design education, artificial intelligence, and virtual reality technology.

Nur Azlina Mohamed Mokmin

Dr. Nur Azlina Mohamed Mokmin is a senior lecturer at the Centre for Instructional Technology and Multimedia, Universiti Sains Malaysia, specializes in Instructional Technology, Virtual Reality, Augmented Reality, and Artificial Intelligence. Holding a master's degree in Instructional Curriculum Development and a Ph.D. in Instructional System, her ongoing research is dedicated to advancing technology for an enriched educational experience for students with disabilities.

Qi Shen

He is an associate professor in the College of Humanities and Arts at Jiangsu Maritime Institute of Technology. He received his engineering degree in industrial design from Yangzhou University, his M.A. degree from the School of Architecture and Design, China University of Mining and Technology, and his Ph.D. degree in design from the Oriental Design University, Taiwan, and is currently conducting research at the Postdoctoral Station of Management Science and Engineering, School of Economics and Management, China University of Mining and Technology. His research involves perceptual engineering, design education, and smart marketing.

Analysis of Artificial Intelligence Policies for Higher Education in Europe

Christian M. Stracke[1]*, Dai Griffiths[2], Dimitra Pappa[3], Senad Bećirović[4], Edda Polz[4], Loredana Perla[5], Annamaria Di Grassi[6], Stefania Massaro[5], Marjana Prifti Skenduli[7], Daniel Burgos[2,8], Veronica Punzo[9], Denise Amram[9], Xenia Ziouvelou[3], Dora Katsamori[3], Sonja Gabriel[10], Nurun Nahar[11], Johannes Schleiss[12], Paul Hollins[11]

[1] University of Bonn, Bonn (Germany)
[2] Research Institute for Innovation & Technology in Education, Universidad Internacional de La Rioja, Logroño (Spain)
[3] Institute of Informatics & Telecommunications, National Centre for Scientific Research "Demokritos" (Greece)
[4] University College of Teacher Education Lower Austria, Baden (Austria)
[5] University of Bari, Bari (Italy)
[6] University of Foggia, Foggia (Italy)
[7] University of New York Tirana, Tirana (Albania)
[8] MIU City University Miami, Miami, Florida (USA)
[9] Scuola Superiore Sant'Anna, Pisa (Italy)
[10] University College for Teacher Education of Christian Churches Vienna/Krems, Krems (Austria)
[11] University of Bolton, Bolton (UK)
[12] Otto von Guericke University Magdeburg, Magdeburg (Germany)

* Corresponding author: stracke@uni-bonn.de

Abstract

This paper analyses 15 AI policies for higher education from eight European countries, drawn from individual universities, from consortia of universities and from government agencies. Based on an overview of current research findings, it focuses the comparison of different aspects among the selected AI policies. The analysis distinguishes between four potential target groups, namely students, teachers, education managers and policy makers. The paper aims at contributing to the further development and improvement of AI policies for higher education through the identification of commonalities and gaps within the existing AI policies. Moreover, it calls for further and in particular evidence-based research to identify the potential and practical impact of AI in higher education and highlights the need to combine AI use in (higher) education with education about AI, often called as AI literacy.

Keywords

I. Introduction

THE need for society to guide the development of Artificial Intelligence (AI) technologies in a way that maximizes their benefits and minimizes their risks is currently driving the development and implementation of AI policies. This is crucial for ensuring that AI systems are designed and used to serve the common good, are compatible with human values and ethical principles and are not misused. AI policies are essential for AI systems to operate fairly, ethically, and transparently according to societal norms and values. They can provide frameworks that enable organizations and citizens to thoughtfully address ethical challenges related to autonomy, bias, explainability, privacy, and accountability, and to ensure that AI systems contribute positively to society [1]. AI systems should not perpetuate or escalate harm or inequality, as in cases of AI-enabled GPT detectors that have frequently misclassified non-native English writing as AI generated, raising concerns about fairness and bias [2]. AI policies are therefore needed to protect individual and public interests but also to encourage innovation in AI tools and applications and to promote cooperation in AI provision and use.

AI policies and regulations exist at different levels and address various stakeholders. In our study we focus on AI policies in higher education as our key research and working field.

The rapid evolution of AI technologies has made available advanced tools for personalized and adaptive learning, data analytics, virtual assistants, and other applications that promise to enhance and disrupt teaching and learning. The uptake of these tools has given rise to a debate in education institutions about readiness, ethics, trust, and the impact and added value of AI, as well as the need for governance, regulation, research and training to cope with the speed and scale at which AI is transforming teaching and learning [3]. At the same time, on-the-ground use of AI by students and teachers is creating a practical need for agreements, policies, and regulatory processes. Alongside this technological potential, specific pedagogical concerns in the university context have emerged, as highlighted by educational research, focusing on personalized learning experiences, inclusion, and accessibility [4]. Experimental applications of AI in various universities have created the need for effective governance to ensure conscious and ethically aligned use, consistent with the traditional roles of educational institutions in providing ethical and responsible quality education [5]. The emerging necessity of a regulatory framework to protect users and the integrity of the higher education system has led to the current development of international ethical standards and regulations regarding AI use.

The examination of AI policies at a European level is a necessary step towards more ethical and cohesive AI frameworks, and more effective use of AI. Similarly, the identification of best practices and ethical guidelines can align the application of AI with broader European values, such as human dignity, privacy, and fairness. The examination of different AI policies at a European level also allows for the identification of the aspects that are addressed and whether advanced AI policies can serve as models to address risks such as biases, privacy issues, and security threats against potential negative impacts. Policies also often address issues of inclusion and accessibility. The comparison between different approaches is an essential step in the society-wide debate that is necessary if AI development is to benefit all segments of society while promoting social equity and avoiding widening the digital divide. Public perception and trust in AI are influenced by how well policies are seen to protect citizens' rights and promote transparency. The comparison of policies helps to identify those approaches that successfully build the public trust and acceptance which is critical for the widespread adoption of AI technologies.

This paper contributes to meeting the need of ethical use of AI and its regulation through the analysis of European AI policies in higher education. The structure of the paper is as follows: section II provides an overview of the background and related work, section III introduces the methodology and selected AI policies, and section IV gives the findings from their analysis. Section V discusses the results and section VI presents the conclusions.

II. Background and Related Work

The recent development of AI-powered technologies has been faster than the development of any previous technology. AI has already been applied in many fields, such as medicine, communication, media, business, and education, enabling tasks to be carried out with greater speed and efficiency, with fascinating and promising results. In this context, educational institutions are both following practice in industry and also developing new technologies [6]. In their teaching they are not only enabling students to adequately use AI-powered tools, but also training them on how to use, improve, and create new AI tools.

The emergence of ChatGPT (Generative Pre-trained Transformer) at the end of November 2022, sparked a debate over the influence of AI-powered technologies on education [7]. AI-driven chatbots such as ChatGPT, Microsoft Bing AI, and Google Gemini are equipped with advanced language models and user-friendly interfaces and these features allow them to engage in human-like conversations and generate original content in response to user prompts [8] [9].

AI encompasses a vast array of technologies, extending far beyond ChatGPT and the various LLMs [10]. These have profound implications for our daily lives, our societies and geopolitics. The intricate web of AI technologies and systems is not a stand-alone entity; rather, it is an integral component of large-scale socio-technical systems [11]. In order to fully comprehend the impact of AI on our future, it is imperative to assess its implications for the redefinition of education and the enhancement of governance capabilities in the coming years [12].

The capabilities of AI offer a number of potential benefits for education [10] [13] [14]. The use of AI-based technologies can provide an opportunity for personalized teaching and learning by diversifying objectives based on each student's abilities, interests, and motivations, with the goal of allowing all students to reach their full potential while neglecting ethical implications [15]. Furthermore, these technologies can enable and improve the individualization of content and teaching methods, promoting flexible and inclusive teaching by adapting methodologies to individual characteristics while pursuing the curriculum's core competences [16]. Evaluating student tasks and generating exercises based on students' current knowledge and expectations, which are assessed by means of placement tests or other forms of testing are just some of the applications of AI-based tools for individualised teaching [17].

In addition to providing assistance and support for students' learning, AI-based technology also offers support to teachers and education managers [18]. AI can assist teachers in updating and designing curricula, creating daily lesson plans and instructional materials, evaluating the student's knowledge, and tracking and reflecting both, students' academic growth and educators' own teaching and professional development. AI-powered tools can also assist university administrators with the analysis, organization, reflection, and use of data [17]. Thus, AI in education could drive a transformation in teaching and learning practices and program development, making it a crucial domain for educational research [19].

AI technologies may be transforming day-to-day educational practice while there is still a lack of evidence-based research and findings on the impact of AI in teaching and learning [7] [14] [20] [21]. However, García-Peñalvo [22] emphasizes the importance of digital transformation for governments, enterprises, and organizations, and warns against unethical use of technology. Hence, in addition to its numerous advantages and benefits, technology, as a whole, can also be employed for harmful objectives; thus, its improper utilization can cause disruption and result in negative consequences [23]. Therefore, it is imperative to implement measures that reduce or mitigate adverse AI outcomes, while simultaneously striving for ongoing enhancement, through the utilization of AI in efficient, meaningful, and ethical methods [24].

Furthermore, it is claimed that the explainability of AI systems is emerging promoting the use of methods that could produce transparent explanations and reasons for decisions made by AI systems. This promised explainability would help to ensure the integrity of the system and at the same time could enable human users to understand, appropriately trust, and effectively manage AI systems [25].

Even though the integration of AI into higher education may be advantageous, or even necessary, it is a topic of great concern for numerous parties because the information provided by AI is not always correct and may have a negative influence on various aspects of students' development. Among the risks raised by an increased use of AI are those associated with data manipulation, intellectual property theft, dealing with sensitive information, harassment, students' social

and emotional development, and ensuring compliance with applicable laws and regulations. Moreover, if not employed appropriately, the implementation of AI technology can have a negative influence on students' capacity for critical thinking, problem-solving, creativity, and ultimately impede their academic success. Such AI-related impacts emerge when students become unduly reliant on automated answers and solutions, neglecting to devote sufficient effort in reflecting them and developing their own solutions to assignments.

Education professionals are becoming increasingly concerned about the use of AI-based tools such as ChatGPT, Jenni AI, Jasper AI, StealthGPT, etc. to write essays and create modules, microlessons and other academic assignments [26]. Bozkurt et al. [10] point out that "we can neither disregard, resist, nor deny the enduring presence of generative AI-driven conversational agents" (p. 201) and that it may "lead students to progress and graduate based on work that is not their own in the traditional sense" (p. 54). Therefore, colleges and universities have to be cautious in regard to "the side practice related to abusive and unethical use and exploitation of data within the learning process" [22, p. 10]. Further, despite the fact that numerous technologies such as Turnitin, ZeroGPT, AI Content Detector, and GPTzero have been developed to identify content that is not human-produced, there are still significant challenges in this area because the precision and reliability of these tools remain in doubt. In fact, they do not provide a solid basis for making appropriate decisions regarding academic integrity [27], which is why many educators and colleges do not accept them.

An additional challenge presented by AI in transforming education is the constraint on social interaction. This is particularly true in an online environment where students are equipped with tools and platforms that enable personalized learning from any place and at any time. In a such a learning environment, students do not have direct engagement with teachers and their peers, which may have a negative impact on their social, psychological, academic, and emotional development. Further, given the widespread use of digital and AI tools and applications, it is nearly impossible to prevent students from being exposed to potentially dangerous content that can negatively impact their emotional, personal, social, and academic growth [28]. As AI-based tools continue to advance, the distinction between authentic content (including text, voice, photo, and video) and content generated by AI technologies is becoming less discernible. Thus, students have already begun to abuse AI technology by creating explicit photographs of their friends in order to harass them [29] [30] [31]. If such practice is not adequately observed, promptly regulated, and carefully prevented, it is foreseeable that instances of unethical AI utilization will escalate and broaden in the future through the improper exploitation of not only peers but also of teachers' and others' voices, videos, and images [32]. Hence, educational institutions must cultivate students' attitudes and consciousness regarding ethical and responsible utilization of AI-based technology, while also implementing measures to prevent misuse, cheating, theft, discrimination, and other examples of improper AI technology use [33].

The ethical, privacy and data-ownership implications arising from the use of AI are substantial concerns. According to Williamson et al. [34], the weight of the available evidence suggests that the current wholesale adoption of unregulated AI applications in schools poses a grave danger to democratic civil society and to individual freedom and liberty. In addition, Pisica et al. [35] point out that "responsibility toward the actions of algorithms, chatbots, and robots, the ethics behind those who create AI and those who operate AI, data privacy, and security are big themes that have been launched in the ethics debate about AI" (p. 4). Thus, considering the aforementioned challenges, numerous governments and educational institutions initially imposed restrictions on students' utilization of AI tools such

as ChatGPT. Likewise, many governments, universities and experts, have expressed concern about the issues of cheating, plagiarism, and the potential detrimental impact of AI use on students' intellectual abilities and cognitive development. Accordingly, experts across the globe have raised their voices for efficient AI regulations, e.g., Birkstedt et al. [36] who state:

> Although public demand for ethical AI continues to grow, if AI technologies are to benefit individuals, then organizations, society, and stakeholders need to be able to trust the technologies and the organizations using them. Academic research should keep pace with the demand and lead discussions on sufficiently broad and practicable AIG (governance) approaches. (p. 160)

This need is also emphasized by Xiao et al. [37] who pointed out that "a comprehensive yet flexible policy could enable faculty and students to reap the benefits of using such technology in the classroom" (p. 4). Thus, efforts are underway to promote the formation of a policy that addresses the ethical and efficient integration of AI into educational practices [38]. Global international organizations such as the Council of Europe, UNESCO, European Commission, the Institute of Electrical and Electronics Engineers [IEEE]), national (UK, China, Japan, USA, etc.) and regional governments, educational and research institutions, enterprises, and other entities make significant efforts to formulate and implement appropriate policies to ensure the responsible and ethical utilization of AI technology in general (e.g., [39] [40]). Below we outline some of the policies of global organizations that are specifically related to the AI use in education (AIED) and significantly influence the policies on national, regional and institutional levels.

UNESCO

To address the issues mentioned above and to support Member States in the use of AIED, UNESCO released a framework for education policymakers aimed at realizing the goals of the 2030 Education Agenda while ensuring the fundamental values of equity and inclusivity in education [41]. With the guidance of policymakers on AI, UNESCO provides educational governance with an overview of basic concepts, methods, and technologies of AI as well as details on the new developments and their impact on teaching and learning using AI in a morally sound, inclusive, and unbiased manner. In addition, suggestions are made on how to learn and work in an AI-driven world and how to improve education and life through the use of AI. The framework also highlights the risks and difficulties of using AI to accomplish Sustainable Development Goal 4 (SDG 4) and provides specific advice on developing strategies and programs that address regional concerns [41].

European Commission

One year after the UNESCO framework had been released, the European Commission and its Directorate-General for Education, Youth, Sport and Culture issued ethical guidelines on the use of AI for educators [42]. Unlike the UNESCO framework which is directed at advising policy makers, the guidelines of the European Commission specifically target support for educators. The guidelines aim to increase teachers' understanding of the potential dangers and assist them in comprehending the power of AI and data usage in the classroom. In this way they enable teachers to utilise AI tools and to deal with them legally, consciously, constructively, and effectively [42].

Council of Europe

In the same year, the Council of Europe (CoE) published a report on the effects of using AI in education seen from the perspectives of Europe's fundamental values [43]. Regarding human rights questions, the focus is on how AI affects children's rights to education, their dignity, autonomy, privacy, data protection, and many more. Further, the report examines how the dominance of commercial AI applications may jeopardize democratic education, how some tools encourage individualism at the expense of social and cooperative

aspects of teaching and learning and how AIED could both strengthen and weaken democratic values. In addition, the legal challenges posed by the use of AI algorithms, such as the use of historical data to grade students, the tracking of learning data and biometric data, are scrutinised [43] [44]. Currently, the Council of Europe is developing with its appointed AI&ED Expert Group an international convention as a specific AI law for education which will be complementary to the EU AI Act [44] [45]. In addition, the CoE is currently working upon a set of actions to facilitate teaching and learning with and about AI such as the AI policy toolbox and a specialised assessment tool for AI in education systems.

EU AI Act

In March 2024, the European Union approved the first comprehensive international framework for limiting the risks of AI: the AI Act [46]. It follows an impact-oriented approach with four categories ranging from unacceptable risk (which is forbidden), high risk (which is regulated), limited risk (which may be taken with some obligations), and minimal risk (which remains unregulated) plus the additional category of general-purpose AI models that was added during later negotiations. Concerning the implementation of AI technology in education, recital 56 of the EU AI Act points out the importance of promoting high-quality digital education and training and allowing all learners and teachers to acquire and share the necessary digital skills and competencies, including media literacy, and critical thinking, to take an active part in the economy, society, and in democratic processes [46]. In accordance with Annex III, AI systems are qualified as 'high risk' if they are intended to a) determine access or admission to educational training, b) be used to evaluate learning outcomes, c) be applied for the purpose of assessing the appropriate level of education that an individual will receive or will be able to access, or d) be used for monitoring and detecting prohibited behaviour of students during tests [46]. In its article 4, the AI Act establishes the obligation of AI literacy for all providers and deployers, requiring them to "take measures to ensure, to their best extent, a sufficient level of AI literacy of their staff and other persons dealing with the operation and use of AI systems on their behalf" [46]. Thus, the role of education and training related to AI systems constitutes an essential pillar of the AI Act with respect to the existing chain through which AI systems are introduced as well as for the possible users, both as individuals and as groups.

AI Policy Implementation in (Higher) Education

Despite attempts to effectively govern the use of AI in (higher) education, there are still criticisms of adopted policies and requirements for them to be updated and improved. Thus, Birkstedt et al. [36] offer that these policies and principles provide little guarantee that they are being applied in reality due to a lack of concreteness and a stronger emphasis on what is happening rather than why what is a surprising statement. Furthermore, regarding the school sector, Williamson et al. [34] "recommend that school leaders pause adoption of AI applications until policymakers have had adequate time to fully educate themselves about AI and to formulate legislation and policy ensuring effective public oversight and control of its school applications" (p. 4). Likewise, the Bipartisan Senate AI Working Group priorities for AI policy in the United States include ensuring enforcement of existing laws for AI, increasing funding for AI innovation, and performing cutting-edge AI research and development, bolstering national security by addressing national security threats, risks, and opportunities for AI, and identifying ways to ensure higher education institutions and companies of all sizes can compete in AI innovation [47]. Hence, even though many global organizations, governments and institutions have adopted AI policies, due to the fast and unpredictable AI technology explosion and development, there is a need to revise existing and develop new policies which should enable effective and safe AI use in higher education.

All co-authors belong to the Network "Ethical Use of AI" (https://ethicalai.ecompetence.eu/) and as a European Network of researchers and teachers from universities, our main interest is in the exploration, analysis and promotion of AI policies for higher education and their potential improvements and practical applications. The European Network is an open and independent initiative run by and for researchers and teachers. Together with all interested colleagues we are meeting monthly without financial interests and without any funding to facilitate better and ethical use of AI in education, to facilitate AI literacy and to serve the society.

Several studies have already analysed and compared AI policies worldwide in different selected countries [48] or from specific rankings [37] [49] or from global institutions [50]. We decided to collect and analyse European AI policies from the eight European countries represented by the co-authors.

III. Methods

This study aims to analyse policies on AI in higher education. This study is purposely conducted at an early stage in AI policy development, which is characterised by a lack of uniformity within and across the different levels of policy making. The domain is characterised by rather disparate initiatives and varying levels of maturity. This study aims to collect and analyse important themes and to highlight emerging directions in current practice, to set the ground for future consolidation and consensus making. Our research questions are "What aspects are addressed in the selected AI policies?" and "How do the policies differ in relation to issuers and target groups."

A. Methods

Our analysis of the AI policies is based on the ethical principles from the AI High Level Expert Group guidelines for trustworthy AI that were published in 2019 after an open consultation with more than 500 contributors [51]. They include human agency and oversight, technical robustness and security, privacy and data governance, transparency, diversity, non-discrimination and fairness, environmental and social welfare, and accountability. Furthermore, they have inspired the EU Regulation on Artificial Intelligence, hereinafter "AI Act" [46] in the definition of grounds of assessment for the compliance of the AI-based systems.

The article performs a formal and content-related analysis of currently published European AI policies for higher education through case study methodology.

For our study, we selected the countries of all co-authors as basis for our research. We have identified and collected AI policies from universities as well as from national public authorities. To keep a balance between the countries, we have selected a national AI policy, if available, and AI policies from universities so that there are not more than four AI policies from each country. To avoid an amount of AI policies that we could not handle and analyse, we have selected and concentrated on AI policies that are from larger universities and already in practice for a longer time. Table I provides an overview of the analysis categories for each AI policy. The country overviews based on the completed tables for all AI policies are published in a separate document that is also published on the website of our European Network with a DOI [52]. The key findings are presented in the section V.

Our analysis framework (Fig.1) takes into consideration the four main target groups who an AI policy can be aimed at: students, teachers (including tutors and lecturers), educational managers (including administrators), and policy makers. In addition, their role (with respect to AI systems/tools), the area of AI application (use cases in higher education) and the scope of an AI policy are focused.

TABLE I. ANALYSIS CATEGORIES FOR THE SELECTED AI POLICIES

Basic information

Name:	[What is the name of the AI policy?]
Issuer:	[Who has authored and issued the AI policy?]
Country:	[In which country is the AI policy developed and issued?]
Educational level:	[Which educational level(s) does the AI policy address? It has to include higher education]
Description:	[What is short description of the AI policy?]
Link, URL, DOI:	[What is the website link, URL or DOI for the AI policy?]

Formal analysis aspects

Application area(s):	[For which application area(s) is the AI policy? Administration, research, teaching, formal examination, self-learning, self-assessment, all?]
Application scope(s):	[For which application scope(s) is the AI policy? Single university, group of universities, all universities in a region/in a country?]
Educational level(s):	[For which educational level(s) is the AI policy? Teachers & students (micro level), Design & teaching (meso level), Organisation & education system (macro level), all levels?]
Policy focus:	[What is the focus of the AI policy? AI use in Education (AIED), education about AI (AI literacy), both = AI and Education (AI&ED)?]
Type:	[Policy or guideline or recommendation?]
Status:	[Published or under development? In the first analysis round, only completed and published AI policies are compared, in a second analysis round, all AI policies (including those that are currently under development) are highlighted in country overview]
Geographical or political	[Is it applied to the whole country, a region, or an organisation?]
Introduction:	[Top-down or bottom-up? Directive of the management or co-creation by teachers, researchers or community?]

Content-related analysis aspects

Ethical considerations:	[Which ethical considerations for developing educational policies are evident in the published documents? Three potential dimensions: meta-ethics, normative ethics, applied ethics]
Ethical principles:	[Which ethical principles are addressed, followed and promoted in the AI policy?]

Students

Role: Users of AI tools

Application: Learning & Research

Guidance: Select and use AI tools correctly

Teachers

Role: Users of AI tools

Application: Teaching & Research

Guidance: Select and use AI tools correctly

Education Managers

Role: Purchasers & Users of ethical AI systems

Application: Leadership & Administration

Guidance: Provide, use and evaluate ethical AI systems

Policy Makers

Role: Supervisors & Regulators of ethical AI systems

Application: Policy planning & Monitoring

Guidance: Control, regulate and safeguard ethical AI systems

Fig. 1. Analysis framework - principal analysis dimensions.

The selected case studies are listed in the following section III.B. And the results of their analysis are presented in section IV.

B. Selected AI Policies

We have selected 15 AI policies from eight countries for our analysis. Table II provides an overview of the selected 15 AI policies and their details.

IV. RESULTS

In the section Results, we present the findings from the analysis of our 15 AI policies. First, we analyse the AI policies according how they are addressing different target groups. Afterwards, we analyse the AI policies according their different status.

A. Analysis of the AI Policies According the Target Groups

We analyse the 15 AI policies according to the four target groups: Students (TG1), Teachers (TG2), Education Managers (TG3) and Policy Makers (TG4) as presented in Table III.

The main criterion for our analysis is the amount of agency for the different target groups according to three potential levels:

1. NR = Not Relevant: the AI policy is not relevant at all for the specific target group,
2. SR = Some Relevance: the AI policy provides some relevance and actions for the specific target group but it is mainly directed to another target group,
3. R = Relevant: the AI policy offers guidance and greater emphasis for the specific target group (sometimes including a vision for the specific target group).

TABLE II. Overview of the Selected AI Policies

ID	Name, Issuer (Country)
1	Guidelines for Dealing with AI in Teaching University of Vienna (Austria) https://phaidra.univie.ac.at/o:1879857
2	Artificial Intelligence Strategy German Federal Government (Germany) https://www.ki-strategie-deutschland.de/files/downloads/Fortschreibung_KI-Strategie_engl.pdf
3	What is Artificial Intelligence (AI)? How can I use AI ethically at university? Network "Ethical use of AI" (Germany) https://doi.org/10.5281/zenodo.10995669
4	Recommendations for the Use of Artificial Intelligence in Academic Performance and Examinations Humboldt University of Berlin (Germany) https://www.hu-berlin.de/de/studium/pservice/empfehlungen_ki_in_pruefungen_hu_2023-09-18.pdf
5	Strategic Programme on Artificial Intelligence (AI) containing 24 policies Italian government (Italy) https://assets.innovazione.gov.it/1637777513-strategic-program-aiweb.pdf
6	LLM Policy of the University of Siena University of Siena (Italy) https://www.unisi.it/sites/default/files/albo_pretorio/allegati/Linee%20guida%20UNISI%20Chat%20GPT_.pdf
7	Hellenic National Strategy for Artificial Intelligence Hellenic Ministry of Digital Governance (Greece) https://digitalstrategy.gov.gr/website/static/website/assets/uploads/digital_strategy.pdf
8	Albania's Digital Agenda and its action plan for 2022-2026 Council of Ministers (Albania) https://www.akshi.gov.al/wp-content/uploads/2022/06/vendim-2022-06-01-370-Agjenda-Digjitale-e-Shqiperise-22-26-dhe-plani-i-veprimit.pdf
9	Generative AI and ChatGPT in the Classroom International Burch University (Bosnia and Herzegovina) https://www.ibu.edu.ba/offices/publications
10	Generative AI in education. Educator and expert views Department for Education (DFE) UK Government (United Kingdom) https://assets.publishing.service.gov.uk/media/65b0c90c160765000d18f74f/DfE_GenAI_in_education_-_Educator_and_expert_views_report.pdf
11	Principles on the use of generative AI tools in Education The Russell Group (UK Research Intensive University Group) (United Kingdom) https://russellgroup.ac.uk/news/new-principles-on-use-of-ai-in-education/
12	Declaration for an ethical use of Artificial Intelligence in Higher Education Universidad Internacional de La Rioja (UNIR) (Spain) https://bit.ly/unir-ia
13	Recommendations for teaching with generative artificial intelligences Carlos III University of Madrid (Spain) http://hdl.handle.net/10016/37989
14	Basic guide on the use of artificial intelligence by teachers and students Autonomous University of Madrid (Spain) https://www.uam.es/uam/media/doc/1606941290988/guia-visual-iagen.pdf
15	Artificial Intelligence in the university University of Granada (Spain) https://ceprud.ugr.es/formacion-tic/inteligencia-artificial

TABLE III. Assignments of AI Policies to Target Groups (TG)

ID	TG1	TG2	TG3	TG4	Comments
1	NR	R	SR	NR	Focused on teacher needs, minimal student focus.
2	NR	NR	NR	R	Emphasizes policy issues, not specific to other groups.
3	NR	R	NR	SR	Addresses systemic teacher issues but not students or managers.
4	SR	SR	SR	NR	Some student focus, less on teachers, involves managers.
5	NR	NR	NR	R	Primarily focused on policy issues.
6	NR	R	NR	NR	Focused mainly on teachers, no student relevance.
7	NR	NR	NR	R	Emphasizes policy maker issues.
8	NR	NR	NR	SR	Similar focus as Policy 7.
9	R	NR	NR	NR	Strong emphasis on student issues.
10	SR	R	SR	R	Addresses both students and teachers, includes management actions.
11	R	R	SR	NR	Greater emphasis on both students and teachers.
12	R	R	SR	SR	Comprehensive approach for students and teachers.
13	R	R	SR	NR	Highlights student and teacher needs.
14	R	R	SR	NR	Strong focus on students and teachers.
15	R	R	SR	NR	Emphasizes student and teacher engagement.

Legend: NR = Not Relevant   -   SR = Some Relevance   -   R = Relevant

In relation to the target group ***Students***, the AI policies selected for analysis can be divided into the following three categories:

1. Several AI policies are not relevant to the student target group (1, 2, 3, 5, 6, 7, 8).
2. Some relevance to students can be found in a few AI policies (4, 10).
3. A larger number of AI policies address the practical issues and these have a greater emphasis on students (9, 11, 12, 13, 14, 15).

In relation to the target group ***Teachers*** *(including Tutors and Lecturers)*, the AI policies selected for analysis can be divided into the following three categories:

1. Some are not relevant to the 'teacher' target group (2, 5, 7, 8, 9).
2. One AI policy addresses systemic issues faced by Universities and thus, it has much less emphasis on the 'teacher' target group (4).
3. The majority of policies address the 'teacher' target group (1, 3, 6, 10, 11, 12, 13, 14, 15).

In relation to the target group ***Education Managers*** *(including Administrators)*, none of the AI policies selected for analysis discuss or plan for the ways that AI can be used to support the activities of academic managers. However, the AI policies identify (to varying degrees) actions which have implications for education managers. In this respect, the AI policies selected for analysis can be divided into the following three categories

1. Some are not relevant to the manager target group (2, 3, 5, 6, 7, 8, 9).
2. The majority of the relevant policies are addressed primarily to teachers and students, but describe some actions or processes which involve education managers or administrators (1, 4, 10, 13, 14, 15).

3. A smaller number address some of the systemic issues raised by Generative Artificial Intelligence (GenAI) for the university or universities, and these imply more actions for academic management and administration (11, 12).

In relation to the target group ***Policy Makers***, the AI policies selected for analysis can be divided into the following four categories:

1. Some are not relevant to the policy maker target group (1, 4, 9, 11, 13, 14, 15).

2. The majority of the relevant policies are addressed primarily to teachers and students, but describe some actions or processes which involve policy makers (3, 8, 12).

3. A smaller number address the systemic issues faced by the government, and these have a greater emphasis on policy makers (2, 5, 7, 10).

4. No AI policy is combining directions for the AI use in combination with a long-term AI vision and declaration that is also relevant for policy makers.

Overall, the majority of the AI policies are focusing mainly the target groups Teachers (9 x relevant, 1 x some relevance) and Students (6 x relevant, 2 x some relevance) while the target groups Policy Makers (4 x relevant, 3 x some relevance) and Education Managers (0 x relevant, 8 x some relevance) are less addressed.

B. Analysis of the AI Policies According the Issuer Groups

There is a distinction to be drawn concerning the level at which the policies were created. On the one hand are individual universities (1, 4, 6, 9, 13, 14, 15), which have responsibility for their own processes. On the other hand, there are consortia of universities (3, 11, 12) and government agencies (2, 5, 7, 8, 10), which are responsible for creating the policy or legal framework within which universities determine their processes.

1. AI Policies From Individual Universities

There is a group of AI policies which provide advice or guidance for teachers and students that has implications for the challenges facing the institution, and for the education managers and administrators who work in it, but without identifying specific challenges or responsibilities. Nevertheless, the fact of creating guidelines is in itself an intervention by management, intended to regulate the use of AI in the delivery of education in the institution.

The clearest case is perhaps **Carlos III University of Madrid** (13), which provides a set of guidelines at the University level, but the guidelines themselves delegate responsibility to those teachers and students, stating that "Any strategy based on limiting, preventing or sanctioning the use of these tools is destined to fail", thereby excluding any direct management of the use of AI. Consequently, the guidelines which it provides are exclusively at the level of individual practice, providing insight which can support teachers and students in the "ethical, correct and effective application" of AI.

Similarly, the **Humboldt University** AI policy (4) focuses exclusively on "the use of generative AI processes that are able to use prompts to generate content that is relevant for answers to exam tasks (e.g., ChatGPT), in exams and in the context of coursework (special work)". The document states that "the use of AI in examinations and coursework should not be generally banned" but it can be restricted or banned for specific purposes. In line with the structure of the University, the detail of such restrictions is delegated to the "faculties and their examination committees to make subject-specific and binding decisions in the knowledge of these recommendations". The policy allows examiners to use AI in the creation of examination documents, but requires them to "observe data protection and copyright regulations in connection with examinations". No guidance

is offered on how this could be achieved, and this is perhaps also delegated to the faculties.

Some other institutions appear to adopt a similar delegation, without being as explicit. For example, the **University of Vienna** AI policy (1) contains a section where high level ethical concerns are outlined, but this is not related to specific challenges for management or administration, and the same is true for legal issues. Similarly, **The Department of Education for England** policy paper on GenAI in education (10) observes that technology has the potential to "reduce workload across the education sector", leaving open the possibility of application to management and administration tasks without any elaboration. It also stresses the existing duty of care of education institutions to avoid access to harmful internet content, and extends this to GenAI without exploring the issues.

There is another subgroup of policies which take a similar approach of delegation to teachers and students, but which do identify some interventions which need to be made by institutional managers and administrators. The provision of training is mentioned by the **Autonomous University of Madrid** (14) and the **University of Siena** (6). The University of Siena also states that sanctions should be applied for the misuse of chatbots by students, and that reflection is required on restructuring and diversifying teaching and assessment. These aspects imply actions by managers and administrators, but the actions are not specified. Intellectual property is discussed, stating that ideas generated by AI cannot be patented, that use of AI must be indicated in papers and student assignments, and that care must be taken "that the use of AI to process personal data does not expose the data subject to any risk or damage." No guidance is provided on how these requirements are to be carried out.

The University of Granada (15) provides guidelines created by the Resource Production Center for the Digital University (CEPRUD). In addition to guidance for teachers and students, these provide three recommendations which articulate the use of AI for the entire university community. Interestingly, two of them relate to the overall university policy on data security, requiring staff to (1) avoid entering sensitive data into a publicly accessible GenAI system, and (2) to use the GenAI tools provided by the University, which protect the privacy of the data that is entered and is not used in the training of the model or transferred to third parties.

2. AI Policies From Consortia of Universities

Three policies are from a consortium of universities (3, 11, 12). The AI policy "What is Artificial Intelligence (AI)? How can I use AI ethically at university?" (3) from the **German Network "Ethical use of AI"** was developed by researchers and teachers from the network that includes more than 40 higher education institutions from Germany, Austria and the Switzerland. Thus, it is unique as it provides the first German guidance for university teachers from themselves. The majority of the network members are responsible for the university-wide further education and training and shared the same demand for a short and easy AI introduction leading to this guidance. Accordingly, it combines a definition of AI in ten statements and a FAQ list with answers and specific practical recommendations how to introduce AI in higher education. It includes requirements for careful reflection and specific regulations to ensure equity as well as the human rights. The main target group are the university teachers but also education managers and policy makers can gain insights concerning issues to be addressed and resolved when considering the use of AI in higher education.

The other two AI policies specifically address some of the systemic issues raised by GenAI. Both of these policies were produced by supra-university bodies, specifically a group of elite universities in the UK (11), and a company which runs multiple universities (12). Thus, the

immediate audience for these documents consists of the managers and administrators the universities which they address, rather than the teachers and learners to who are the principal audience for the individual university policies which we have examined. It is therefore unsurprising that these two AI policies imply more actions to be taken by education managers and administrators. However, in neither of them there is mention of the use of GenAI in the execution of university administration or management. **The Russell Group** (11) of elite UK universities has published principles on the use of GenAI tools, affirming their members' intentions. These imply some actions for managers and administrators, for example the application of institution-wide policies, and the need to consider university-wide subscription to GenAI services, and the provision of training. **The Universidad Internacional de La Rioja (UNIR)** (12) belongs to the holding company Proeduca Education Group which includes nine universities (with 130,000 students) and a VET institution worldwide. Like the Russell Group, the UNIR declaration of principles for the use of GenAI has implications for management and administration, including the need for training, but in this case with a strong focus on responsible use of AI, including ethical procedures, data privacy, human control of AI, transparency and traceability. Uniquely among the documents we have surveyed, the principles include an internal audit and monitoring system, which would require action by management and administration. The principles have been applied in a Guide for the Responsible Use of Generative AI in Research Tasks [53], and may be applied to other areas of university activity as and when UNIR identifies a need.

3. AI Policies From Government Agencies

The remaining AI policies are developed by a government agency (2, 5, 7, 8, 10).

The "Artificial Intelligence Strategy of the **German Federal Government**" (2) provides political guidelines for all sectors in general but not specifically for the educational sector while education is mainly addressed as (vocational) education and training for other sectors.

In addition, the "National Strategic Programme on Artificial Intelligence" (5) by the **Italian government** focuses on all sectors while AIED is mainly seen as task for higher education. In addition, it asks for AI education at all levels but without precise tasks and realizations.

The "Digital Transformation Guide 2020-2025" (7) by the **Greek government** with its 422 pages is the longest policy, but only a specific section concerns education. This section discusses potential challenges and covers mainly managerial aspects for future changes.

"Albania's Digital Agenda and its action plan for 2022-2026" by the **Council of Ministers from Albania** (8) also focuses more on other sectors than education and mainly on technological issues. AI is explicitly mentioned but the implementation and risks of AI are only discussed related cloud computing, cybersecurity and managerial issues.

The policy paper "Generative artificial intelligence (AI) in education" (10) by the **UK government** only addresses school and college education and consequently its relevance for our research focus on higher education is limited.

V. Discussion

In this section, we reflect on the results from our country descriptions and the analysis of the AI policies related to the four target groups.

First of all, an important main finding is that we could not easily identify AI policies in education and not many universities have developed and published their own guidelines. The adoption of national, European and international AI policies remains fragmented, with only a few universities having adopted them and the targets set by these policies not always being consistent. It is evident that government policies on AI in the educational sector are not always present, and there is a need for greater awareness of the governance of AI tools. The concern is that, as during the COVID-19 pandemic [54] [55], we move in an unstructured manner and without a common and shared vision, compromising the quality and inclusiveness of formal education in the face of these global changes, and risking the generation of inequalities, a lack of equity in access and injustice. Therefore, we hope that future harmonisation of AI policies at the regional (European) level will ensure equal opportunities for future European citizens.

The second key finding is that the role of students varies greatly in these AI policies for higher education, particularly in relation to GenAI tools. A critical distinction emerges: some AI policies position students as relatively *passive* recipients of instruction, while others promote a more *proactive* role. The former view students as beneficiaries of the actions of higher education institutions and their teachers, while the latter considers student agency in the use of AI.

Certain AI policies appear to adopt a **passive student role**. These assume that students' engagement with AI depends on the actions of external actors, universities and educators or should be guided by them. The focus therefore lies on empowering and supporting students through initiatives directed towards universities and teachers, namely actions aimed at communicating limitations and ethical considerations surrounding AI use, particularly generative AI (GenAI), namely improving critical reasoning skills, and training students on safe and responsible use of GenAI tools. At the same time AI use is regarded as the means for: (a) enhancing teaching practices and student learning experiences; and (b) promoting the development of future-focused skills in students.

Table IV features indicative statements extracted from guidance documents that frame student engagement with GenAI as contingent upon actions taken by external actors (universities and educators). This guidance is directed towards teachers and higher education institutions.

Conversely, other guidance documents, particularly those stemming from Spanish universities, promote a **proactive student model**. These documents present GenAI as a "learning partner" and "writing partner" for students, encouraging responsible and ethically aware utilization. This approach is evident in statements such as:

- "...teach students how to use emerging technologies, such as generative AI, safely and appropriately..."
- "...support students, particularly young pupils, to identify and use appropriate resources..."

This framing emphasizes student agency and initiative in harnessing GenAI for their educational benefit.

Overall, the passive model positions students as beneficiaries of actions undertaken by teachers and institutions, while the proactive model empowers students to utilize GenAI as an active learning tool.

Some of the policies examined also include specific recommendations for teachers. These emphasise the need to obtain consensus for the use of AI systems, develop university-wide ethical guidelines, and cooperate with other institutions to improve practices in the use of AI.

The implementation of AI in higher education offers a promising avenue for enhancing the academic management and delivery of education. Nevertheless, it is of the utmost importance to address ethical challenges and ensure that the use of these technologies is accountable, transparent and respects the rights of individuals, namely students and teachers. The application of AI in the field of academic

TABLE IV. Statement Examples Framing Passive Student Role

Statement (no. of AI policy)	AI for T/L	AI for Skills	Safe AI
Lecturers should communicate to students early on whether and, if so, within what limits, the use of AI is permitted in their examinations and coursework (4)			X
The education sector needs to: … teach students how to use emerging technologies, such as generative AI, safely and appropriately (10)			X
The education system should: support students, particularly young pupils, to identify and use appropriate resources to support their ongoing education; encourage effective use of age-appropriate resources (which, in some instances, may include generative AI); prevent over-reliance on a limited number of tools or resources (10)			X
Our universities wish to ensure that generative AI tools can be used to … enhance teaching practices and student learning experiences, ensure students develop skills for the future within an ethical framework (11)	X	X	
Staff should be equipped to support students to use generative AI tools effectively and appropriately in their learning experience (11)			X
Improve critical reasoning skills and prepare students for the real-world applications of the generative AI technologies they will encounter (11)		X	
Employees, students, teachers and other professionals will be trained to follow responsible practices in the use, distribution, dissemination and production of AI-based technologies and services, consistent with the group's ethical standards (12)			X

Legend: T/L = Teaching & Learning

administration is regarded as a potential means of enhancing the management and delivery of education, to support students and teachers. The various AI policies analysed provide guidance on the selection of the most appropriate AI tools and the effective and safe use of these tools. On the other hand, the development of ethically sound systems is mostly concerned and addressed with data collection for setting up and improving models rather than what the models do. Nevertheless, several ethical challenges associated with the use of AI are highlighted, including the necessity to maintain human accountability in decisions made with the aid of AI, to ensure the accessibility and reliability of tools for all users, to guarantee that decisions are transparent and fair, and to protect the privacy and security of data processed by AI systems.

Only a few AI policies call for the combination of AI use in education (AIED) and education about AI (AI literacy), namely the AI policies by the German and Greek governments (2, 7) and the guidance by the German Network "Ethical Use of AI" (3).

While it could be expected that the majority of AI policies would mainly target university teachers and managers, it is surprising that they do not contain many recommendations and requests directed to policy makers.

Through the analysis and comparison of regional AI policies, we investigated whether or not universities in the selected countries had adopted university-level policies and identified which aspects were essential to consider when comparing the selected AI policies. The objective is to provide an analysis and reflection exercise that might inform and facilitate the development of contextualised frameworks for the ethical and responsible adoption of AI technologies.

Thus, this comparative study and its results represent only a preliminary investigation for further research. To gain a more profound comprehension and a comprehensive global and analytical perspective, it is imperative to conduct more extensive and longitudinal research to collect, compare and evaluate all policies.

One approach could be the development and continuous evaluation and improvement of an index for the AI readiness, use and impact in regions and countries regarding higher education. For the generic field of AI, such a matrix is proposed by the annual Government AI Readiness Index released from Oxford Insights [36]. Although its foundations are not transparent and thus, it could be critiqued, also for potential biases, this Index is a measure for assessing how ready governments are to adopt and manage AI technologies. It consists of three Pillars: Technology Sector (focusing supply, innovation and human capital), Data & Infrastructure (focusing on their availability and high quality) and Government that is most important for our analysis. Aspects of relevance to education that are part of the Government Pillar include governments' strategic vision for AI development and governance, supporting regulation for governance and ethics, and development of internal digital capacity in terms of skills and practices for adapting to change. In particular, the Governance and Ethics Dimension of the Government Pillar examines aspects such as Data protection and privacy legislation, Cybersecurity Regulatory Quality and Accountability, while the Vision Dimension investigates national AI strategies [56]. Currently, the Government AI Readiness Index is often used as basis for research even though it does not cover all the specifics of education and lacks an evidence-based research approach [57]. Thus, it cannot be easily applied to identify the potential and practical impact of AI in (higher) education. Developing a dedicated AI Readiness Index for education can help shed light on how AI adoption is currently progressing in (higher) education and contribute to the goal of a future where AI is applied responsibly and effectively to enhance research, teaching, and learning.

Finally, we emphasise that there is an urgent need to combine AI use in (higher) education with education about AI, often referred to as AI literacy. This is necessary to ensure that all educational stakeholders and target groups (students, teachers, education managers and policy makers) can understand and reflect AI and are aware of the potential opportunities and risks of AI use in (higher) education.

VI. Conclusion

We have argued that AI, unlike traditional information and communication technologies (ICT), presents unique ethical and social challenges, including data security, algorithm transparency, social impact and educational quality, and ethical responsibility. There is no consensus on the ethical aspects of AI as a technological practice, and therefore, the AI development is guided solely by the principles of those who create and deploy it. Consequently, the ethical aspects of policies and declarations are dealing with personal and individual positions put forward not only, but in particular to the people involved. The situation is made more complex by the connection between regulatory adaptations and the rapid development of AI. Regulatory frameworks that apply at all levels in higher education, both nationally and internationally, as well as institutional guidelines,

are therefore necessary to guarantee that AI in higher education is implemented ethically, transparently, and with respect for human rights. An overarching framework can provide a theoretical approach to the topic, while practical guidelines can offer contextualized answers to questions clustered by topic, sector, target group, etc.

The development of regulations and ethical frameworks for the responsible utilization of AI in universities is presently at a nascent stage. While governments across Europe are taking significant steps to establish regulatory standards for AI use in public sectors, comprehensive national policies for the responsible and ethical use of AI in Education are not available. Presently only sporadic, mostly bottom-up initiatives exist to develop regulations and ethical frameworks for the responsible and ethical use of AI in higher education. Most universities and academic institutions are in the early stages of implementing their own structured approaches to AI ethics and governance. Regulations are always one step behind practice, and so it is inevitable, and perhaps desirable, that the institutions push forward in their policies on and practice with AI, and so ensure that regulations are designed in relation to the real impact of the technology.

The present research provides the basis for developing comprehensive and relevant guidelines for the ethical use of AI in higher education to ensure that all stakeholders are able to deal responsibly with the complexities of AI. From this perspective, we have highlighted several critical aspects for creating effective guidelines on the ethical and responsible use of AI in higher education. These guidelines should address different target groups, define roles in AI interaction, cover diverse application areas, and provide a clear scope for their guidance.

Our findings call for further and in particular evidence-based research to identify the potential and practical impact of AI in higher education. There is an urgent need to always combine AI use in (higher) education with education about AI, often called AI literacy, to ensure that all stakeholders and target groups (students, teachers, education managers and policy makers) are aware of the potential opportunities and risks of AI use in (higher) education. In the final analysis, AI is not ethical nor moral; people are. AI policies in education should aspire to supporting organizations and citizens in fulfilling this responsibility.

References

[1] L. Floridi, "*The Ethics of Artificial Intelligence,*" Oxford University Press, 2020.

[2] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, & J. Zou, "GPT detectors are biased against non-native English writers," *Patterns 4*, July 14, 2023, https://doi.org/10.1016/J.patter.2023.100779

[3] M. Bond, H. Khosravi, M. De Laat, N. Bergdahl, V. Negrea, E. Oxley, ..., & G. Siemens, "A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour," *International Journal of Educational Technology in Higher Education*, vol. 21, no. 4, 2024, https://doi.org/10.1186/s41239-023-00436-z

[4] L. Perla, & V. Vinci, "Enhancing Authentic Assessment in Higher Education: leveraging Digital Transformation and Artificial Intelligence," In *AIxEDU 2023. High-performance Artificial Intelligence Systems in Education, 3605* (pp. 1-7). CEUR-WS, 2023.

[5] C. M. Stracke, B. Bohr, S. Gabriel, N. Galla, M.Hofmann, H. Karolyi, ..., & G. Stroot, "*Ethical Use of Artificial Intelligence (AI) in Higher Education – a Handout,*" 2024a, https://doi.org/10.5281/zenodo.10995669

[6] S. Bećirović, "Examining learning management system success: A multiperspective framework," *Education and Information Technologies*, vol. 29, pp. 11675–11699, 2023, https://doi.org/10.1007/s10639-023-12308-0

[7] C. M. Stracke, I.-A. Chounta, & W. Holmes, "Global trends in scientific debates on trustworthy and ethical Artificial Intelligence and education," *Artificial Intelligence in Education. CCIS*, vol. 2150, pp. 254–262, 2024, https://doi.org/10.1007/978-3-031-64315-6_21

[8] A. Bozkurt, "Generative artificial intelligence (AI) powered conversational educational agents: The inevitable paradigm shift," *Asian Journal of Distance Education*, vol. 18, no. 1, pp. 198–204, 2023, https://doi.org/10.5281/zenodo.7716416

[9] M. Liu, L. J. Zhang, & C. Biebricher, "Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing," *Computers & Education*, vol. 211, no. 104977, 2024, https://doi.org/10.1016/j.compedu.2023.104977

[10] A. Bozkurt, J. Xiao, S. Lambert, A. Pazurek, H. Crompton, S. Koseoglu, ..., & P. Jandrić, "Speculative futures on ChatGPT and generative artificial intelligence (AI): A collective reflection from the educational landscape," *Asian Journal of Distance Education*, vol. 18, no. 1, pp. 53–130, 2023, https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/709/394

[11] R. Huang, A. Tlili, L. Xu, Y. Chen, L. Zheng, A. H. Saleh Metwally, ..., & C. J. Bonk, "Educational futures of intelligent synergies between humans, digital twins, avatars, and robots - the iSTAR framework," *Journal of Applied Learning & Teaching,* vol. 6, no. 2, pp. 28-43, 2023, https://doi.org/10.37074/jalt.2023.6.2.33

[12] C. Bentley, C. Aicardi, S. Poveda, L. Magela Cunha, D. Kohan Marzagao, R. Glover, ... & O. Acar, *A Framework for Responsible AI Education: A Working Paper,* 2023, August 17, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4544010

[13] A. Bozkurt, J. Xiao, R. Farrow, J. Y. H. Bai, C. Nerantzi, S. Moore, ..., & T. I. Asino, "The Manifesto for Teaching and Learning in a Time of Generative AI: A Critical Collective Stance to Better Navigate the Future," *Open Praxis*, vol. 16, no. 4, pp. 487–513, 2024, https://openpraxis.org/articles/10.55982/openpraxis.16.4.777

[14] A. Tlili, M. A. Adarkwah, C. K. Lo, A. Bozkurt, D. Burgos, C. J. Bonk, ..., & R. Huang, "Taming the Monster: How can Open Education Promote the Effective and Safe use of Generative AI in Education?," *Journal of Learning for Development*, vol. 11, no. 3, pp. 398-413, 2024, https://doi.org/10.56059/jl4d.v11i3.1657

[15] L. Tang, & Y.-S. Su, „Ethical Implications and Principles of Using Artificial Intelligence Models in the Classroom: A Systematic Literature Review", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 5, pp. 25-36, 2024, doi: 10.9781/ijimai.2024.02.010

[16] D. T. K. Ng, J. K. L. Leung, K. W. S. Chu, & M. S. Qiao, "AI Literacy: Definition, Teaching, Evaluation and Ethical Issues," *Proceedings of the Association for Information Science and Technology*, vol. 58, no. 1, 2021, https://doi.org/10.1002/pra2.487

[17] S. Bećirović, & B. Mattoš, "Artificial Intelligence in the Transformation of Higher Education: Threats, Promises and Implementation Strategies," *Digital Transformation in Higher Education, Part A* ,pp. 23–43, 2024, https://doi.org/10.1108/978-1-83549-480-620241002

[18] M. Alier, F.-J. García-Peñalvo, & J. D. Camba, „Generative Artificial Intelligence in Education: From Deceptive to Disruptive", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 5, pp. 5-14, 2024, doi: 10.9781/ijimai.2024.02.011

[19] T. K. F. Chiu, Q. Xia, X. Zhou, C. S. Chai, & M. Cheng, "Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education," *Computers and Education: Artificial Intelligence*, vol. 4, no. 100118, 2023, https://doi.org/10.1016/j.caeai.2022.100118

[20] W. Holmes, & I. Tuomi, "State of the art and practice in AI in education," *European Journal of Education,* vol. 57, pp. 542–570, 2022, https://doi.org/10.1111/ejed.12533

[21] C. M. Stracke, I.-A. Chounta, W. Holmes, A. Tlili, & A. Bozkurt, "A standardised PRISMA-based protocol for systematic reviews of the scientific literature on Artificial Intelligence and education (AI&ED)," *Journal of Applied Learning and Teaching*, vol. 6, no. 2, pp. 64-70, 2023, https://doi.org/10.37074/jalt.2023.6.2.38

[22] F. J. García-Peñalvo, "Avoiding the Dark Side of Digital Transformation in Teaching. An Institutional Reference Framework for eLearning in Higher Education," *Sustainability*, vol. 13, no. 4, 2021, https://doi.org/10.3390/su13042023

[23] S. Bećirović, "*Digital Pedagogy: The Use of Digital Technologies in Contemporary Education,*" Springer Nature, 2023, https://doi.org/10.1007/978-981-99-0444-0

[24] C. Adams, P. Pente, G. Lemermeyer, & G. Rockwell, "Artificial Intelligence Ethics Guidelines for K-12 Education: A Review of the Global Landscape," In: *Artificial Intelligence in Education. LNCS, 12749*, 2021, https://doi.org/10.1007/978-3-030-78270-2_4

[25] H. Khosravi, S. B. Shum, G. Chen, C. Conati, Y. S. Tsai, J. Kay, ... & D. Gašević, "Explainable artificial intelligence in education," *Computers and Education: Artificial Intelligence,* vol. 3, no. 100074, 2022, https://doi.org/10.1016/j.caeai.2022.100074

[26] Intelligent, "*Nearly 1 in 3 College Students Have Used ChatGPT on Written Assignments,*" 2023, January 23, https://www.intelligent.com/nearly-1-in-3-college-students-have-used-chatgpt-on-written-assignments

[27] M. Perkins, J. Roe, B. H. Vu, D. Postma, D. Hickerson, J. McGaughran, & H. Q. Khuat, "Simple techniques to bypass GenAI text detectors: Implications for inclusive education," *International Journal of Educational Technology in Higher Education*, vol. 21, no. 1, pp. 1–25, 2024, https://doi.org/10.1186/s41239-024-00487-w

[28] M. Madaio, S. L. Blodgett, E. Mayfield, & E. Dixon-Román, "Beyond "fairness": Structural (in)justice lenses on AI for education," In *The Ethics of Artificial Intelligence in Education.* Routledge, 2022.

[29] M. Klee, "An AI Nightmare Has Arrived for Twitter—And the FBI," *Rolling Stone*, 2023, October 17, https://www.rollingstone.com/culture/culture-features/ai-explicit-material-twitter-big-tech-1234855484/

[30] C. Long, "Children are using AI to bully their peers using sexually explicit generated images, eSafety commissioner says," *ABC News*, 2023, August 15, https://www.abc.net.au/news/2023-08-16/esafety-commisioner-warns-ai-safety-must-improve/102733628

[31] J. Revell, "*All Students Will Be Taught How to Use AI Next Year (By Humans),*" 2023, October 6, https://thelatch.com.au/ai-in-schools

[32] D. Aggarwal, D. Sharma, & A. B. Saxena, "Adoption of Artificial Intelligence (AI) For Development of Smart Education as the Future of a Sustainable Education System," *Journal of Artificial Intelligence, Machine Learning and Neural Network*, vol. 3, no. 6, 2023, https://doi.org/10.55529/jaimlnn.36.23.28

[33] S. Akgun, & C. Greenhow, "Artificial intelligence in education: Addressing ethical challenges in K-12 settings," *AI and Ethics*, vol. 2, no. 3, 2022, https://doi.org/10.1007/s43681-021-00096-7

[34] B. Williamson, A. Molnar, & F. Boninger, *Time for a Pause: Without Effective Public Oversight, AI in Schools Will Do More Harm Than Good,* 2024, March 5, https://nepc.colorado.edu/publication/ai

[35] A. I. Pisica, T. Edu, R. M. Zaharia, & R. Zaharia, "Implementing Artificial Intelligence in Higher Education: Pros and Cons from the Perspectives of Academics," *Societies*, vol. 13, no. 5, 2023, https://doi.org/10.3390/soc13050118

[36] T. Birkstedt, M. Minkkinen, A. Tandon, & M. Mäntymäki, "AI governance: Themes, knowledge gaps and future agendas," *Internet Research*, vol. 33, no. 7, pp. 133–167, 2023, https://doi.org/10.1108/INTR-01-2022-0042

[37] P. Xiao, Y. Chen, & W. Bao, "*Waiting, Banning, and Embracing: An Empirical Analysis of Adapting Policies for Generative AI in Higher Education,*" SSRN Scholarly Paper 4458269, 2023, https://doi.org/10.2139/ssrn.4458269

[38] C. M. Stracke, „Artificial Intelligence and Education: Ethical Questions and Guidelines for Their Relations Based on Human Rights, Democracy and the Rule of Law," *Radical Solutions for Artificial Intelligence and Digital Transformation in Education*, pp. 97-107, 2024, https://doi.org/10.1007/978-981-97-8638-1_7

[39] UK government's Central Digital and Data Office, "*Generative AI Framework for HM Government,*" 2024, https://www.gov.uk/government/publications/generative-ai-framework-for-hm

[40] United Nations Educational, Scientific and Cultural Organization (UNESCO), *Recommendation on the Ethics of Artificial Intelligence,* 2021, https://unesdoc.unesco.org/ark:/48223/pf0000381137

[41] F. Miao, W. Holmes, R. Huang, R., H. & Zhang, *AI and education: Guidance for policy-makers,* UNESCO Digital Library, 2021, https://unesdoc.unesco.org/ark:48223/pf0000376709

[42] European Commission, Directorate-General for Education, Youth, Sport and Culture, *Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators,* 2022, https://data.europa.eu/doi/10.2766/153756

[43] W. Holmes, J. Persson, I.-A. Chounta, B. Wasson, & V. Dimitrova, *Artificial intelligence and education—A critical view through the lens of human rights, democracy and the rule of law,* Council of Europe, 2022, https://rm.coe.int/artificial-intelligence-and-education-a-critical-view-through-the-lens/1680a886bd

[44] C. M. Stracke, I.-A. Chounta, V. Dimitrova, B. Havinga, & W. Holmes, "Ethical AI and Education: The need for international regulation to foster human rights, democracy and the rule of law," *Artificial Intelligence in Education. CCIS*, vol. 2151, pp. 439-445, 2024c, https://doi.org/10.1007/978-3-031-64312-5_55

[45] W. Holmes, C. M. Stracke, I.-A. Chounta, D. Allen, D., Baten, V. Dimitrova, ..., & B. Wasson, "AI and Education. A View Through the Lens of Human Rights, Democracy and the Rule of Law. Legal and Organizational Requirements," *Artificial Intelligence in Education. CCIS, vol. 1831*, pp. 79-84, 2023, https://doi.org/10.1007/978-3-031-36336-8_12

[46] European Union, *AI Act (Artificial Intelligence Act). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence,* 2024, http://data.europa.eu/eli/reg/2024/1689/oj

[47] C. Schumer, M. Rounds, M. Heinrich, T. & Young, T., *A Roadmap for Artificial Intelligence Policy in the U.S. Senate,* 2024, May, https://www.schumer.senate.gov/imo/media/doc/Roadmap_Electronic1.32pm.pdf

[48] N. McDonald, A. Johri, A. Ali, & A. Hingle, "Generative Artificial Intelligence in Higher Education: Evidence from an Analysis of Institutional Policies and Guidelines," *ArXiv*, 2024, https://doi.org/10.48550/arXiv.2402.01659

[49] A. Dabis, & C. Csáki, "AI and ethics: Investigating the first policy responses of higher education institutions to the challenge of generative AI," *Humanities and Social Sciences Communications*, vol. 11, no. 1006, 2024, https://doi.org/10.1057/s41599-024-03526-z

[50] C. K. Y. Chan, "A comprehensive AI policy education framework for university teaching and learning," *International Journal of Educational Technology in Higher Education,* vol. 20, no. 38, 2023, https://doi.org/10.1186/s41239-023-00408-3

[51] High-Level Expert Group (HLEG) on AI, *Ethics Guidelines for Trustworthy AI*, 2019, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[52] C. M. Stracke, D. Griffiths, D. Pappa, S. Bećirović, E. Polz, L. Perla, A. Di Grassi, S. Massaro, M. Prifti Skenduli, D. Burgos, V. Punzo, D. Amram, X. Ziouvelou, D. Katsamori, S. Gabriel, N. Nahar, J. Schleiss, & P. Hollins, *Analysis of Artificial Intelligence Policies for Higher Education in Europe: Country reports*, 2024, https://doi.org/10.5281/zenodo.14811950

[53] Universidad Internacional de La Rioja (UNIR), *Guide to the Responsible Use of Generative Artificial Intelligence in Research Tasks,* Proeduca, 2024, https://doi.org/10.5281/zenodo.11609187

[54] C. M. Stracke, D. Burgos, G. Santos-Hermosa, A. Bozkurt, R. C. Sharma, C. Swiatek, ..., & V. Truong, "Responding to the initial challenge of COVID-19 pandemic: Analysis of international responses and impact in school and higher education," *Sustainability,* vol. 14, no. 3, 1876, 2022a, https://doi.org/10.3390/su14031876

[55] C. M. Stracke, R. C. Sharma, A. Bozkurt, D. Burgos, C. Swiatek, A. Inamorato dos Santos, ..., & V. Truong, "Impact of COVID-19 on formal education: An international review on practices and potentials of Open Education at a distance," *The International Review of Research in Open and Distributed Learning*, vol. 23, no. 4, pp. 1-18, 2022b, https://doi.org/10.19173/irrodl.v23i4.6120

[56] Oxford Insights, *The Government AI Readiness Index,* 2023, https://oxfordinsights.com/ai-readiness/ai-readiness-index

[57] S. Nzobonimpa, & J.-F. Savard, "Ready but irresponsible? Analysis of the Government Artificial Intelligence Readiness Index," *Policy & Internet,* vol. 15, no. 397–414, 2023, https://doi.org/10.1002/poi3.351

Christian M. Stracke

Dr. Christian M. Stracke developed interdisciplinary expertise in research, education and management while leading international large-scale projects (budgets over 50 million euros and up to 200+ researchers). As globally recognized expert, innovator and appointed ICDE Chair in OER, he published 200+ scientific publications on his research fields: Open Education, Artificial Intelligence, Technology-Enhanced Learning, Competence Building, Impact Assessment and Educational Policies (https://orcid.org/0000-0001-9656-8298). He consults many educational institutions, international ministries and global organisations including UNESCO, Council of Europe, European Commission and European Parliament. As a member of the AI&ED Expert Group appointed by the Council of Europe, he works on the international law for regulating AI use in education and on the AI literacy recommendation. Christian is coordinator for the cloud strategy at the German University of Bonn and establishing the research lab at the scientific University IT and Data Center Bonn. He co-founded the UNESCO Unitwin Network on Open Education (UNOE) and holds a PhD in Economics and Informatics (University of Duisburg-Essen, 2014) and a Magister Artium (M.A.) in Educational Sciences (University of Bonn, 2000) as well as professorships at the East China Normal University in Shanghai and the Korean National Open University in Seoul. He serves as editor of peer-reviewed indexed journals and as chair of international conferences. He facilitated and developed global learning standards as elected Chair of ISO, IEC and CEN committees and provided more than 100 keynotes worldwide. And he is Founder and Director of eLC, the European Institute for Learning, Innovation and Cooperation. His website: http://www.opening-up.education.

Dai Griffiths

In 2009 Dai was appointed Professor of Educational Cybernetics at the University of Bolton. In 2015 Dai became a professor in the Department of Education at the University of Bolton, where he has taken on the task of coordinating the PhD and Doctor of Education programmes, as well as supervising PhD candidates, as well as lecturing on research methodologies and global trends in education. Today, he is Professor and Senior researcher at the Research Institute for Innovation & Technology in Education of the Universidad Internacional de La Rioja in Spain.

Dimitra Pappa

Dr. Dimitra Pappa possesses a degree in Electrical Engineering, with a specialization in Telecommunications, from the National Technical University of Athens (NTUA). Additionally, she holds an MBA from the Hellenic Open University and a PhD from the University of Surrey in the United Kingdom. Dr. Pappa has been actively involved in numerous European and national research and development initiatives in the fields of Information Systems for Technology Enhanced Learning (TEL), eLearning, and Knowledge Management. Her contributions to these projects have included serving as project coordinator, research supervisor, project leader, and team member. Her notable project involvements include, but are not limited to, Positive Learn, Health Cascade, Phaetons, Advance, Cloudledge, CRe-AM, OEI2, OpenScout, PROLIX, PROLEARN, and COCAL. Furthermore, she has authored several articles that have been published in peer-reviewed international scientific journals and conferences, demonstrating her commitment to advancing knowledge in her field.

Senad Bećirović

Dr. Senad Bećirović is a passionate and respected full professor at the University College of Teacher Education Lower Austria. With over 20 years of dedicated teaching experience across a wide range of educational settings, he has earned a reputation as a trusted and inspiring academic leader. Professor Bećirović is the author of four influential books and 65 insightful scientific articles. His work has been recognized by renowned publishers such as Springer, Sage, Wiley, and Taylor & Francis, and is widely referenced in respected databases like WoS, Scopus, Q1, and Q2. His research has made a lasting impact on the fields of artificial intelligence in education, digital technologies in teaching and learning, and intercultural education. Beyond his writing, Dr. Bećirović is a sought-after speaker who has shared his expertise at international conferences, both as a keynote and regular speaker. His groundbreaking research continues to inspire and shape educational practices around the world, leaving a profound mark on the future of education.

Edda Polz

Edda Polz has been Vice Rector for Research and University Development at the University College of Teacher Education Lower Austria since 2022. Prior to this, she had been a lecturer in English, school law and educational science since 2014 and was Chair of the University College prior to becoming Vice Rector. She holds a bachelor's degree in primary school teaching (BEd), a master's degree (Mag. iur.) in law, a doctorate in education and communication (PhD) as well as a master's degree (MEd) and a doctoral degree (Dr.) in educational science. Her research interests include teaching English as a foreign language, competency-based education, inquiry-based learning, giftedness, lifelong learning and AI. She regularly publishes articles on these topics in influential books and internationally recognised journals.

Loredana Perla

Loredana Perla is Dean of the Department of Education, Psychology and Communication Sciences and Full Professor of Didactics and Special Education at the University of Bari. She is the scientific coordinator of the Ministerial Commission responsible for defining of guidelines for the new curricula in primary and secondary schools. She is the Italian representative of ISATT (International Study Association on Teaching and Teachers) and of Reseau Ideki (Information - Innovation - Didactics - Documentation - Education - Knowledge - Engineering). She is the director of the "e-learning and health promotion" research unit at CITEL (Telemedicine Centre) of University of Bari. Together with Ettore Felisatti, she coordinated the Anvur working group on "Recognition and enhancement of teaching skills in university teaching", which led to the drafting of the relevant guidelines in 2023. She is the coordinator of Uniba's TLC for Faculty Development and, as the Rector's delegate for teacher training, she is the scientific director of DIDASCO, a service centre for the teaching and professional development of teachers. She founded the Lediel (Leadership, Empowerment and Digital Innovation in Education) Associate Doctorate at Uniba for the training of academic leaders. She also chairs the Scientific Ethics Committee of the research project "Guidelines Verse-Uniba for the use of AI in teaching", for the drafting of guidelines for the co-design and use of AI tools. She is the director of the "Didattizzazione" series published by FrancoAngeli. She is the national coordinator of the SIRD Teacher Training Observatory and the author of over 400 publications.

Annamaria Di Grassi

Annamaria Di Grassi is a PhD student in Learning Sciences and Digital Technologies at the University of Foggia, Italy, with a research base at the Department of Educational Sciences, Psychology, Communication at the University of Bari. She is a member of the Educational Experimentation Laboratory and the DiDasco research group, directed by Prof. Perla, at the same department. Annamaria was a part of the Scientific Ethics Committee for the research project "Guidelines Verse-Uniba for the use of AI in Education", for the drafting of guidelines for the co-design and use of AI tools. She is also a teacher of law and economics in secondary schools and is involved in citizenship and media education. Annamaria has been the coordinator of digital innovation and prevention of bullying and cyberbullying in her school. Her research focuses on the ethical dimensions of transparency and accountability of artificial intelligence, with particular attention to the concept of explainability. She investigates how the integration of AI into academic curricula can foster the development of critical thinking and responsibility, promoting a conscious and ethical use of technologies to empower new generations to face the challenges of the future.

Stefania Massaro

Stefania Massaro is an Associate Professor of Didactics and Special Pedagogy at the Department of Education, Psychology, and Communication Sciences at the University of Bari in Italy. She teaches Special Pedagogy and School-Based Play Methodology in undergraduate and postgraduate programs. She is a founding member of the CITEL University Research Centre in Telemedicine, specifically within the E-Health Education and Wellbeing research unit. Additionally, she is member of ISATT (International Study Association on Teaching and Teachers) and ASDUNI (Italian Association for the Promotion and Development of Didactics, Learning, and Teaching in Universities). Her research involvement includes participation in national projects such as AmICA (Intelligent Holistic Care for Active Ageing in Indoor and Outdoor Ecosystems) and SISTER (Social Robots to Support the Biopsychosocial Frailty of Seniors at Home for the Promotion of Active Ageing). She collaborates with U!REKA European University Alliance for the SHIFT project activities. She is the author of Metamorphosis of Democracy: Principles and Methodologies.

Marjana Prifti Skënduli

Marjana Prifti Skënduli received her Ph.D. in Computer Science from the University of New York Tirana and is currently a Postdoctoral Researcher in Quantum Machine Learning at the Università degli studi di Bari Aldo Moro, Italy. She holds the position of Assistant Professor and Director of Graduate Programs at the University of New York Tirana (currently on sabbatical leave). Dr. Skënduli is a member of the Council of Europe Expert Group on Artificial Intelligence and Education, and serves as a Red Team Network Member and Policy Researcher for large language model provider companies. Her research spans Artificial Intelligence, Machine Learning, Quantum Machine Learning, Data Mining, and Natural Language Processing, alongside a strong interest in AI applications in Education. In 2023, she founded AI Albania, a non-profit organization dedicated to fostering AI research, ethical AI development, and the cultivation of a sustainable AI ecosystem in Albania. Additionally, Dr. Skënduli serves as the Scholarships Committee Chair for ACM-W Global, where she advocates for the advancement and representation of women in computing.

Daniel Burgos

Daniel Burgos is a full professor of Technology for education and communication and vice-rector for international research at the Universidad Internacional de La Rioja (UNIR). He holds a UNESCO Chair on eLearning. He is the Director of the Research Institute for Innovation and Technology in Education (UNIR iTED, http://ited.unir.net). Also, he is the President of MIU City University Miami, USA. He has implemented more than 80 European and worldwide R&D projects and published more than 270 scientific papers, 60 books and special issues, and 12 patents. He is a full professor at An-Najah National University (Palestine), an adjunct professor at Universidad Nacional de Colombia (UNAL, Colombia), an extraordinary professor at North-West University (South Africa), a visiting professor at the China National Engineering Research Center for Cyberlearning Intelligent Technology (CIT Research Center, China); and a research fellow at INTI International University (Malaysia). He works as a consultant for the United Nations (UNECE), the United Nations University (UNU-FLORES), ICESCO, the European Commission and Parliament, and the Russian Academy of Science. He holds 13 PhD degrees and doctorates, including Computer Science and Education. ORCID: 0000-0003-0498-1101.

Veronica Punzo

Veronica Punzo is attorney at the Macerata Bar and a certified economics and law teacher as well as a secondary school specialised teacher for special educational needs in secondary school. In collaboration with the National PhD Course in "Artificial Intelligence for Society" at the University of Pisa, Veronica Punzo is member of the LIDER-Lab at Sant'Anna School of Advanced Studies, Pisa, and teaching fellows in Special Pedagogy at the University of Macerata (IT). Dealing with the governance and regulation of personal and non-personal data and ethical-legal consulting, her research area focuses on the domain of the implementability of AI-based educational resources and regulatory compliance with a focus on generative AI impact on dignity, equality and the enhancement and protection of fundamental rights.

Denise Amram

Denise Amram (PhD in Law) is a Senior Assistant Professor of Comparative Private Law at LIDER Lab - DIRPOLIS Institute and affiliate at the L'EMbeDS Department of Excellence at Scuola Superiore Sant'Anna (Pisa, Italy). She obtained the National qualification as an Associate Professor in 2020. Coordinator of the Permanent Observatory on Personal Injury Damages, of the research lines ETHOS (EThics and law witH and fOr reSearch) and Family Law within the LIDER Lab, she participates, also with leading roles, to several research projects and infrastructures funded under national and EU programs, either dealing with personal and non-personal data governance and regulation or serving as a legal-ethical advisor. Denise has been recently appointed within a pool of experts providing consultancy services for the Council of Europe on Children's Rights and Artificial Intelligence and Education domains and as an evaluator of projects proposals for the EU Commission. Denise enriched her experience undertaking teaching / research activities both in Italy and abroad, including Université Panthéon-Assas and Université Panthéon-Sorbonne, Utrecht University, University College Dublin, University of Malta, Columbia Law School. Denise is attorney at the State Bar of Pisa, enrolled in the special list of professors and researchers, mediator in civil and commercial matters, and ISO certified UNI 11697:2017 as Data Protection Officer. Steering committee member for top ranked law journals (eg Opinio Juris in Comparatione, and GenIUS), she authored / co-authored ~130 publications, including a monograph book. Her research interests include dignity and fundamental rights enhancement and protection-especially in the digital dimension .

Xenia Ziouvelou

Xenia Ziouvelou holds a PhD in Internet Economics and Strategy (scholarship awarded) from Lancaster University, UK. She is currently a Research Associate at the Institute of Informatics and Telecommunications (IIT) at NCSR "Demokritos" and leads the AI Politeia Lab, an interdisciplinary research group exploring the ethical, socio-technical, and policy implications of AI-driven innovation; aiming to ensure responsible, trustworthy, and democratized AI for all and for good. Dr. Ziouvelou is also an external innovation expert for the European Commission and a member of the Expert Group on AI and Education at the Council of Europe. Xenia also serves on the Scientific Committee of AI4People, Europe's first global forum on AI's social impact. Additionally, she represents Greece as a national delegate to the EU AI Board's Steering Group on General Purpose AI and co-chairs the AI Focus Group of the European DIGITAL SME Alliance, the largest network of ICT SMEs in Europe. Xenia's research lies at the intersection of AI, innovation, strategy, policy, and ethics, focusing on responsible AI-driven innovation across domains including education. Passionate about democratizing AI sustainably, she advocates for a human- and value-centric approach to AI-driven innovation, that promotes social progress, economic growth, and environmental sustainability.

Dora Katsamori

Dora Katsamori is an Associate Post-doctoral Researcher in the field of citizenship education and inclusion of the Institute of Informatics and Telecommunications (II&T) at the National Centre for Scientific Research 'Demokritos' (NCSR 'D', Greece). She holds a BA in Political Science and Public Administration from the University of Athens, where she was specialized in the field of International and European Studies, a Master degree in Social Discrimination, Migration and Citizenship and a PhD in the field of citizenship and social disadvantaged groups from the University of the Peloponnese. She is also a certified adult educator of non-formal education. Lately, her field of research concerns AI & Education as well as explainable AI systems as a member of European working groups. She has a long experience as project coordinator and researcher in European and national funded research projects and as educator in the field of civic and political education. She teaches courses in the field of sociology of education at the University of Patras and she is involved in the field of refugee and prison education in Greece.

Sonja Gabriel

Dr. Sonja Gabriel teaches and conducts research as a university professor of media education and media didactics at Private University College of Teacher Education of Christian Churches Austria. After studying German and English studies at the University of Vienna, she completed two master programs (Educational Media, Applied Game Science) and wrote her dissertation in the field of educational science at PH Weingarten. Her current work and research focus on the use of generative AI in education with special focus on learning with large language models and ethical challenges, digital game-based learning, and teaching and learning in the digital age. In her lectures and publications, she particularly emphasizes the ethical use of digital media in educational contexts.

Nurun Nahar

Nurun Nahar is an Assistant Teaching Professor based at the Greater Manchester Business School, University of Greater Manchester. Nurun's responsibilities include overseeing and advising on technology enhanced learning initiatives to enhance pedagogical practices within her department. Nurun is a published scholar and has presented her research work widely at several international conferences alongside invited guest talks on the topics of digital literacy, pedagogical partnerships, use of generative AI and technology enhanced learning in Higher Education.

Johannes Schleiss

Johannes Schleiss is a researcher at the Artificial Intelligence Lab of the Otto-von-Guericke University Magdeburg, Germany. His work focuses on developing innovative and applied educational concepts for Artificial Intelligence (AI) and integrating AI technologies into education. In addition to his research role, he is an associated research fellow at the AI Campus, a leading learning platform for AI, and future scout for generative AI in higher education, funded by the Reinhard Frank Foundation and the Stifterverband. He has also collaborated with the student Think&Do-Tank of the Higher Education Forum on Digitalisation in Germany, where he has contributed to vision building and strengthened student voices in AI. He holds a Master of Science in Digital Engineering from Otto-von-Guericke University Magdeburg (2020) and a Bachelor of Engineering from the Nuremberg Institute of Technology, Germany (2017).

Paul Hollins

Paul is a professor of cultural research and Knowledge Exchange at the University of Bolton in the UK and is a Fellow of the Cybernetic Society. His PhD was focused on exploring the efficacy of immersive environments in formal educational settings, awarded by the University of Bolton in the UK. He also has postgraduate degrees in Education, a Master of Science in learning and teaching with Information & communication Technologies awarded by Leeds Becket University in the UK and a Master of Business Administration awarded by Leeds Business School. His first degree was in Management studies. Paul's research interests are varied but lie in the intersection of technology, learning and digital Games. He has also published extensively in Music. He has directed and participated in several research projects in these spaces over the last three decades including the EU framework funded Learning Interoperability Framework Europe (LIFE), Ten Competence and Realising Applied Games Ecosystem (RAGE) and is currently leading the University of Bolton's participation in the EU Horizon funded Games Realising Effective & Affective Transformation (GREAT) research and innovation project. Paul has published over one hundred and thirty academic outputs and has acted as external examiner for several institutions.