Facial Expressions in Virtual Reality Based-Education: Understanding Recognition Approaches and Their Integration for Immersive Experiences

Anass Touima [1,*] and Mohamed Moughit [1,2]

[1] Laboratory Science and Technology for Engineering (LaSTI), National School of Applied Sciences-Khouribga, Sultan Moulay Slimane University, Morocco

[2] Laboratory Artificial Intelligence, Modeling and Computational Engineering (AIMCE), ENSAM Casablanca, Hassan II University, Casablanca, Morocco

Email: anass.touima@usms.ac.ma (A.T.); mohamed.moughit@usms.ac.ma (M.M.)

*Corresponding author

***Abstract*—Integrating facial expressions into Virtual Reality (VR) for education is hindered by the cost and technical limitations of current Facial Expression Recognition (FER) systems, impacting accessibility and the enrichment of remote learning. Our research aimed to develop and assess a cost-effective, webcam-based FER system for real-time replication of a teacher's facial expressions onto a VR avatar, to enhance emotional interactivity and pedagogical effectiveness in distance education. A Convolutional Neural Network (CNN)-Deep Neural Network (DNN) deep learning model with Correlation-based Feature Selection (CFS) was developed for FER and integrated into a Unity-based VR classroom, using OpenFace for landmark detection from webcam input. Accuracy was validated on benchmark datasets (CK+, JAFFE, OULU CASIA-VIS), followed by an empirical study with 65 instructors. The FER model achieved high accuracy (e.g., 100% on CK+), and our VR application successfully mapped expressions in real-time. Instructors reported improved emotional communication (74%) and engagement (72%), with the system's affordability (approx. $1200/user) being a key advantage, though adoption barriers and occasional misclassifications were noted. In conclusion, an affordable, webcam-based FER system can enhance VR education by improving emotional interactivity. While promising, addressing real-world robustness, facial occlusion by VR headsets, and user acceptance is crucial for wider deployment. Future work includes predicting occluded facial features and multimodal emotion detection.**

***Keywords*—virtual reality, E-learning, facial expressions recognition**

I. Introduction

Virtual Reality (VR) technology has proven to be an innovative technology in education, providing interactive, three-dimensional environments that go beyond the constraints of traditional pedagogies. Through the emulation of real-world situations and experiential learning, VR has proven to increase engagement, retention of knowledge, and motivation among learners in various fields of education [1]. In this context, facial expressions are an important part of communication, not only indicating emotional states but also serving as vital cues for social communication, understanding, and feedback during the learning process. Therefore, the use of Facial Expression Recognition (FER) systems in VR-oriented learning environments is an innovative direction for developing emotionally sensitive and adaptive learning experiences.

Recent research highlighted the increasing importance of FER systems in promoting student emotional engagement and measuring affect states in online learning environments. For example, Aly [2] proposed an improved FER model founded upon ResNet-50 with Convolutional Block Attention Modules (CBAM) and Temporal Convolutional Networks (TCNs), with the goal of overcoming real-time emotion detection issues. This model performed well with benchmark dataset RAF-DB, FER2013, Cohn-Kanade (CK+), and KDEF-FER, with good performance in identifying broad ranges of emotional expressions. In the same vein, Zhang *et al*. [3] considered an optimized model of MobileNet V2 as an option for real-time emotion detection in the VR environment, emphasizing balancing computing power with the precision of identification. These advancements are part of an even more general trend toward designing light, real-time FER models that can be incorporated into VR platforms.

Despite these developments, there remain many issues with the effective use of FER systems in virtual classrooms. One of the biggest issues is the inconsistency

Manuscript received March 27, 2025; revised April 15, 2025; accepted June 10, 2025; published November 25, 2025.

and unreliability of recognizing emotions in unstructured and changing environments, particularly with low-cost or consumer-grade devices. Experiments have demonstrated that lighting variability, partial occlusion, and facial morphology variation among individuals can severely affect recognition performance [4]. Additionally, current systems are not able to effectively detect subtle or mixed emotions, which would restrict their use in delicate educational interactions where confusion, frustration, or disengagement are critical to pedagogical effectiveness [3].

Another concern is the privacy and ethical issues of perpetual facial surveillance. Some methods try to address these concerns through avoiding data storage or anonymizing the input, yet many commercial solutions are opaque about how user data is processed and protected [4]. Most of today's VR headsets do not support real-time facial tracking or need costly, specialized sensors, which limits accessibility and prevents adoption in underprivileged educational institutions [3].

To overcome these constraints, we introduce an improved FER system for VR environments. The solution utilizes a dual-modal deep learning framework that blends Convolutional Neural Networks (CNNs) with fully connected Deep Neural Networks (DNNs), capturing both facial shape and texture characteristics of expressions. It makes possible precise real-time mapping of emotions to avatars without the need for specialized VR-computing infrastructure, ensuring greater accessibility and cost-effectiveness.

The principal contributions of this work are as follows:

- Development of an efficient, light-weight CNN-DNN model for facial expression classification with high accuracy but with compatibility with common webcams and customer-grade devices.
- Integration of the suggested FER system into an evolving virtual reality classroom environment, allowing real-time facial expressions of the users to be projected onto avatars, hence creating increased emotional interactivity for distant learning.
- Implementation of Correlation-based Feature Selection (CFS) to lower the facial descriptor dimensionality and enhance classification efficacy while overcoming overfitting and enhancing generalizability over various datasets.
- Extensive testing over various benchmarking datasets (CK+, JAFFE, OULU CASIA-VIS) and cross-dataset verification to measure robustness and flexibility of the proposed technique.
- Empirical testing with instructors at the university to assess the pedagogical efficacy and usability of the VR classroom app, with insights into technical performance, adoption, and future areas for improvement.

Recent developments from 2023 to 2025 further solidify the need for Facial Expression Recognition (FER) to be integrated into immersive and learning environments. Aly [2] proposed an improved FER model with ResNet- 50 augmented with Convolutional Block Attention Modules (CBAM), which reached over 95% accuracy on benchmarking data such as RAF-DB and FER2013, while resolving privacy issues with non-invasive data handling methods. This model not only enhanced feature extraction with an emphasis on prominent facial regions but highlighted ethical issues of increasing concern in affective computing applications. Meanwhile, Zhang *et al*. [3] explored the use of optimized MobileNet V2 models for real-time emotion detection in virtual reality, with the need for balancing the cost of computations with high-level recognition accuracy, particularly when utilized with consumer-level devices such as the Meta Quest Pro [3]. More recently, in 2025, there have been initial attempts to address domain adaptation tasks, with efforts aimed at enhancing model generalizability over different populations and lighting environments [5]. These models make use of both spatial attention and temporal modeling in order to detect subtle emotional fluctuations, which align with pedagogical objectives where confusion, engagement, or frustration need to be understood in order to implement adaptive learning. In addition, Aly *et al*. [4] introduced an attention-driven FER model suited for online learning platforms, allowing real-time monitoring of the emotional engagement of students without sacrificing responsiveness in the system. This work further encourages the interest in affective computing in education, with support from large-scale meta-analytic evidence documenting the positive effects of emotionally responsive pedagogical tools on motivation and learning among learners [1] which continues to gain further developments in early 2025.

Building from these advancements, our research helps continue the endeavor to address the disparity between theoretical study and real-world application of FER systems in virtual reality avatars for education. By prioritizing cost-effectiveness, scalability, and real-world relevance, we seek to offer an effective solution that facilitates emotionally intelligent, interactive learning environments available to more people. VR has become a technology tool, in settings providing immersive and interactive experiences that go beyond traditional learning boundaries. VR immerses users in virtual environments where they may directly observe and control items [1]. As teachers incorporate VR into their teaching methods, recognizing the significance of expressions in this context is crucial. Facial expressions play a role as communication signals aiding emotional expressions, social engagements and under-standing during in-person interactions. Using VR technology in education accurately captures, interprets, and reproduces expressions and boosts the involvement and nurturing of the socio-emotional learning and enhances educational outcomes.

Facial expressions are central to human interactions and rather than using verbal words, speak volumes about a person's feelings, intentions and attitudes. A Study have also shown that individuals with higher emotional awareness are better equipped to handle social interactions, leading to deeper connections and reduced misunderstandings [6]. In learning situations, teachers use

verbal and non-verbal communication to show empathy, motivation, and enthusiasm, which enhances interactive learning environments. Similarly, students' facial expressions give teachers insight into their level of understanding, engagement, and emotional well-being.

The realistic experience of virtual reality enhances the importance of facial expressions in educational contexts, by creating scenarios and human interactions, as well as, VR environments provide teachers and students with the opportunity to engage in authentic and emotionally impactful learning experiences. However, obstacles such as inaccuracies in tracking delay, responsiveness, and the challenge of achieving attractive visuals pose significant challenges to the full use of facial expressions in VR based education [7, 8]. To overcome these obstacles, we need to make progress in hardware and software innovations, and research collaborations across all fields.

II. Background

A. Virtual Reality in Education

Virtual Reality (VR) is one of the major breakthroughs in the education sector, revolutionizing the old classroom by immersing students in simulations and interactive environments. It further encourages experiential learning by allowing learners to interact actively with educational content. In VR, students get a chance to engage with historical events, explore scientific principles, and develop practical competencies in a secure and regulated environment. The new technology accommodates learning preferences through personalized experiences that suit needs and tastes. Further, VR technology is transcendent in that learners can access material from wherever they are [9]. Increased adoption of VR in environments makes students have better under-standing, memory retention, and motivation to learn, resulting in a more engaging and successful academic journey.

B. Importance of Facial Expressions in Education

Facial expressions play a role in environments, acting as a key element, for successful communication, understanding and social engagement between teachers and students. By using expressions, educators' express emotions, goals and perspectives enhancing the learning journey, with nonverbal signals that support spoken guidance. Studies emphasize the role of expressions, in promoting engagement and motivation in students [10]. When educators convey encouragement, empathy, or enthusiasm through their expressions, they build connections with students and create a conducive learning atmosphere. Moreover, facial expressions help in understanding cues and emotions, enabling teachers to assess students' understanding, interest and emotional well-being [11]. Furthermore, integrating expressions, into materials and multimedia presentations improve learning outcomes by adding context and reinforcing verbal explanations [12]. Therefore, recognizing, and interpreting expressions do not support effective communication and feedback, but also enhance socio emotional growth and cultural awareness among learners [13]. Teachers need to understand how important facial expressions are, in education and use them effectively to enhance students' involvement, understanding and emotional development.

III. Related Work

Recent research highlighted the increasing value of Facial Expression Recognition (FER) systems in creating emotional engagement and tracking students' affective states in online learning. Towards this goal, Aly [2] proposed an improved FER framework that combines ResNet50, the Convolutional Block Attention Module (CBAM), and Temporal Convolutional Networks (TCNs), aimed at resolving issues with real-time emotion identification. The model shows robust performance with benchmark datasets commonly utilized, such as RAF-DB, FER2013, CK+, and KDEF-FER, with accuracy values above 91%. By virtue of spatial attention principles and temporal modeling methods, the system can effectively overcome adversaries such as light variation, partial occlusion, facial morphology differences among individuals, and the temporal variability of emotional facial expressions. Some of its benefits are the delivery of real-time emotional feedback to instructors, the potential for pedagogical interventions customized to each student, and increased student engagement through evidence-based insights. However, the use of such systems poses privacy concerns and issues of informed consent, especially in educational situations with children or vulnerable individuals. More, there are limitations in accurately identifying ambiguous or culturally sensitive facial emotions, while technical shortcomings such as reliance on good video input and computing power may decelerate popularization, particularly in schools with weak technological infrastructure.

There have been recent attempts to integrate FER systems into virtual reality environments to improve interactivity and emotional engagement of users. Specifically, Zhang *et al*. [3] have considered the application of an optimization of the MobileNet V2 model as a real-time emotion recognizer in virtual reality, with the Meta Quest Pro headset and Unity engine. The study reveals that emotions like "Happiness", "Sadness", and "Neutral" were accurately recognized, indicating that light-weight deep models can effectively operate in environments with limited resources. Most importantly, the system provided dynamic remapping of users' emotions onto avatars without storing or transmitting personal facial information, mitigating privacy issues commonly linked to affective computing technologies. These findings demonstrate the potential applications of FER-based systems in areas from immersive education to therapy, where emotional feedback would effectively enhance user experience and interactivity.

In spite of these positive findings, the study also found there were certain limitations. Lower recognition rates for emotions like "Anger" and "Fear" can likely be attributed to overlapping facial features and lack of training samples, and indicates more general issues of dataset imbalance and model generality. In addition, the relatively low resolution

and positioning of the front-facing cameras in modern VR headsets restrict the quality of facial expressions captured, particularly in changing lighting environments or partial occlusions. Such tangible restrictions could compromise the accurate identification of subtle or mixed emotional states. Moreover, exclusive use of facial expressions ignores other necessary modes of emotional display, such as body language or speech, indicating the necessity of further holistic, multi-modal strategies for recognizing emotions in VR environments.

Aly *et al*. [4] suggested an improved Facial Expression Recognition System (FERS) suited for online learning environments to track the emotional states of students and monitor their engagement in real time. The suggested model utilizes an improved ResNet-50 architecture with Convolutional Block Attention Modules (CBAM), which helps in filtering out irrelevant visual noise and emphasizing essential facial areas like the eyes, eyebrows, and mouth. This helps improve the model's discrimination power to recognize faint emotional expressions. On benchmarking with data sets like RAF-DB and FER2013, the system showed high accuracy, demonstrating that it is effective in detecting varied emotional states. Due to this, this system helps in near real-time tracking of learners' emotions during virtual classes, which proves to be very helpful for instructors to make adjustments to their instruction and enhance student learning.

Despite these positive findings, several limitations need to be considered. One issue is the model's generalizability to other demographics—facial expressions differ significantly among cultures, genders, and age groups, which could compromise the performance of the system in various educational environments. Second, while the model performs robustly in controlled environments, deployment in real-world scenarios may be impeded by technological constraints such as minimal available computing resources or poor video quality inputs, particularly in under-resourced institutions. Ethical issues surrounding continuous facial tracking, especially with concerns for data privacy, user consent, and possible misappropriation of sensitive information, are also noteworthy. Third, exclusive use of facial features may not provide complete or accurate interpretations of student engagement, which makes additional modalities such as voice or behavioral tracking necessary to create more comprehensive emotion recognition systems.

IV. Facial Expressions Recognition Approaches

A. Geometric Approach

This approach consists in measuring the change in facial expressions using geo-metric descriptors. These are based on distances calculated between a few fiducial points. The system first detects the face; then it discloses 49 characteristic points on the face; and from a few points, distances are calculated to form our descriptors. Subsequently, the proposed descriptors are introduced to a classifier to train it to classify facial expressions into emotions of the discrete category.

1) Detection of the face

The first step after acquiring 2D images, is face detection (Fig. 1). This step is necessary to limit the area of interest of the data processing for a good representation of the features.

To carry out this step, we used the Viola-Jones Algorithm [14], utilizes a set of filters called HAAR descriptors for the extraction of features, and for the classification, a set of cascaded classifiers.

Fig. 1. Face detection with the Viola and Jones algorithm [13].

2) Detection of fiducial points

After detecting the face, all the images used by the system were resized to a resolution of 256×256 pixels to keep the same size. Then, the fiducial points were detected on the new 256×256-pixel image, which contained only the facial area. To locate the points, our choice fell on the Supervised Descent Method (SDM) [15]. This method is fast, robust, and suitable for real-time data processing.

It is trained to detect 49 feature points on the face as illustrated in Fig. 2. These points represent the positions of the facial components. The points detected are represented in a two-dimensional space, and they are distributed on the face as follows: 5 characteristic points for each eyebrow, 18 points for the lips, 9 points for the nose, and 6 points for each eye.

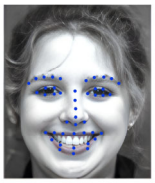

Fig. 2. The 49 points detected by SDM [15].

3) Extraction distances

Once the fiducial points are detected, the next step is to extract the descriptors representing the facial expression. The descriptors that we proposed are based on the calculation of the distances between some characteristic points among the 49 points located on the face. The calculated distances are shown in Fig. 3, they represent distances between relevant areas of the face, considering the area of the nose which is generally neglected by several studies, such as:

- D1: the distance between the eye and the eyebrow;
- D2: the distance between the eyelashes of an eye;
- D3: the distance between the mouth and the eye;
- D4: the distance between the mouth and the nose;
- D5: opening of the mouth;
- D6: the distance between the upper lip and lower lip of the mouth.

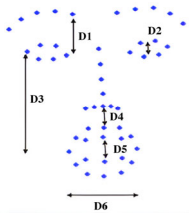

Fig. 3. The descriptive distances of the face used as features.

4) Classification of facial expressions

In this part, we present the different methods used in classification step in addition to the process followed for static images and dynamic images.

a) Classification And Regression Tree (CART)

It is a basic machine-learning algorithm that serves two purposes: classification and regression [16]. It works by recursively partitioning the feature space into different regions based on feature values, with each partition associated with a class label in classification or a numerical value in regression. CART creates a binary tree where each internal node represents a feature-based decision and each leaf node corresponds to a predicted class label or numerical value. By iteratively selecting feature splits that minimize impurity or variance, CART creates interpretable decision tree models, suitable for a variety of tasks. Its simplicity, flexibility, and ability to handle both categorical and continuous functions make it a popular choice for a variety of machine learning applications.

b) Network artificial neurons

A Multilayer Perceptron (MLP) is a basic type of Artificial Neural Network (ANN) consisting of interconnected layers of neurons. They process information through feedforward propagation, where each neuron applies a weighted sum and activation function to its input. By adjusting these weights using backpropagation during training, MLPs can learn complex patterns and relationships in the data, making them a versatile model for tasks such as classification, regression, and pattern recognition in various fields [17].

c) The Support Vector Machine (SVM)

It is a versatile supervised learning algorithm used for classification and regression tasks. It works by finding the best hyperplane that best distinguishes different classes in the input space. By maximizing the margin between support vectors, SVM achieves robustness and generalization to new data. Support vector machines can handle complex data sets and non-linear relationships through kernel techniques, and are widely used in various fields due to their effective-ness and efficiency in pattern recognition and decision-making tasks [18].

5) Databases

We can cite two databases of known facial expressions: the extended Cohn-Kanade (CK+) database [19], and the JAFFE database [20]:

Cohn-Kanade (CK+): Developed by researchers at the University of Pittsburgh, the image consists of more than 5000 images of facial expressions depicting a variety of emotions including happiness, sadness, anger, surprise, disgust and fear (Fig. 4). The images were captured by 210 subjects with varying intensities and changes in facial expressions. Due to its scale and diverse expressive capabilities, CK+ is particularly compelling for use in training and evaluating facial expression recognition algorithms.

The Japanese Female Facial Expression (JAFFE) [20]: On the other hand, this database contains 213 grayscale images of facial expressions of 10 Japanese female models. The database focuses on six basic emotions: happiness, sadness, surprise, anger, disgust, and fear (Fig. 5). Each image is labeled with a corresponding emotion label, making it suitable for training and testing facial expression recognition systems.

Fig. 4. Examples of images from the CK+ database [19].

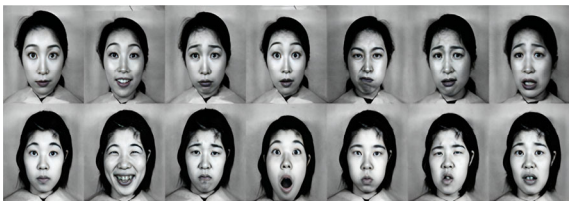

Fig. 5. Examples of images from the JAFFE database [20].

6) Experimental protocol

Our study evaluated facial expression recognition methods using the CK+ and JAFFE datasets. CK+, as detailed in the Table I, includes 118 subjects performing 7 expressions (6 basic+contempt), divided into 92 training and 26 test subjects to ensure subject-independent results. JAFFE contains static images of 10 Japanese women, split into 8 training and 2 test subjects. Three classifiers (CART, MLP, SVM) were tested with Euclidean, Minkowski, and Manhattan distance descriptors on both static images (single emotion) and dynamic sequences (neutral → emotional). Configurations included 6 classes (excluding contempt/neutrality), 7 classes (with contempt), and 7* classes (with neutrality) to assess diverse scenarios.

TABLE I. NUMBER OF EXPRESSIONS USED IN EACH EXPERIMENT FOR THE CK+ AND JAFFE DATABASES

Databases	**CK+**			**JAFFE**	
Number of classes	6	7	7*	6	7*
Anger	45	45	45	30	30
Disgust	59	59	59	29	29
Joy	69	69	69	31	31
Fear	25	25	25	32	32
Sadness	28	28	28	31	31
Surprise	83	83	83	30	30
Contempt	-	18	-	-	-
Neutrality	-	-	118	-	30
Total	309	327	427	183	213

Manhattan distance outperformed other metrics, achieving higher classification rates—e.g., 78.57% accuracy with SVM on JAFFE. Dynamic data improved CK+ performance (e.g., 93.26% with SVM) by capturing expression evolution (Fig. 6), while static images yielded better results on JAFFE (Fig. 7), likely due to its limited subject diversity. SVM and MLP consistently surpassed CART, despite CART's simplicity. Including neutrality (7* classes) reduced CK+ accuracy (e.g., 77.52% with CART), and contempt (7 classes) faced limitations from its small sample size (18 instances).

Our study confirms Manhattan distance and advanced classifiers (SVM/MLP) enable robust, subject-independent emotion recognition. Static images proved sufficient for competitive performance, simplifying systems for constrained datasets like JAFFE. These findings support automated solutions classifying isolated images without prior references (neutral state/sequences). Future work could explore deeper architectures or complex descriptors to overcome geometric approach limitations, enhancing scalability and accuracy across diverse populations and expressions.

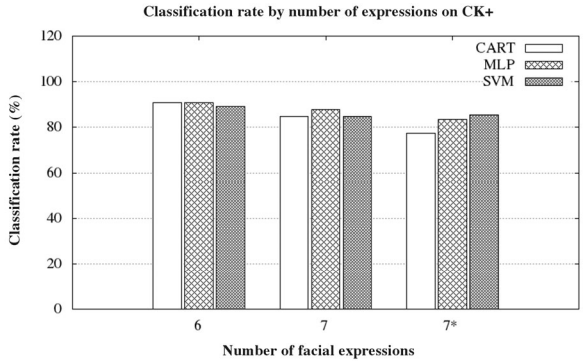

Fig. 6. Comparison of classification rates between CART, MLP, and SVM classifiers on static images with Manhattan descriptors on the CK+ database.

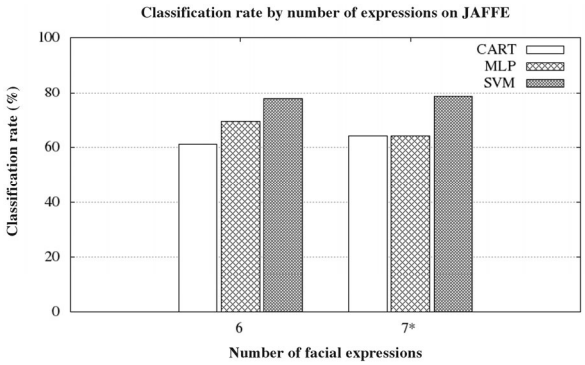

Fig. 7. Comparison of classification rates between CART, MLP, and SVM classifiers on static images with Manhattan descriptors on the JAFFE database.

B. Learning by Selection of Relevant Features

1) Description of the proposed approach

The main steps in emotion recognition from facial expressions are face detection, feature extraction, and classification. It is necessary to detect the face first. Then, traits or features that better describe the emotion must be discovered, and finally, these features must be grouped into basic emotions. The second step is the extraction of features, which is where the problem comes from. Identifying and employing the best facial features for classification is crucial.

Fig. 8 illustrates the approach proposed, which is based on automatic learning by selection of relevant features. Pre-processing and detection of the face precede the extraction of the morphological identifiers. Then, a step is taken to select the most relevant descriptors. Finally, the classification step is performed using only the selected descriptors. Fig. 8 illustrates the proposed approach, which is based on automatic learning by selecting the relevant features. The face is first detected and pre-processed, then geometric descriptors are extracted. After that, a step of selecting the descriptors is carried out to keep only the most relevant ones. Only the selected descriptors are used for the classification phase.

Fig. 8. Steps of selection of relevant features approach.

2) Representation and classification of facial expressions

a) Features extraction

Next, we will extract the set of descriptors that represent the facial expression. The descriptors presented in the previous section were based on the calculation of six distances between 12 characteristic points among the 49 points located with SDM. The deformations of the facial components are covered by these distances, which are chosen manually. There may be distances that are more descriptive than the ones we have chosen. Therefore, we suggest here to calculate all the possible distances between each pair of points among the 49 points located on the face in order to measure all the possible deformations (Fig. 9). We obtain in total $C^2 = 1176$ distances.

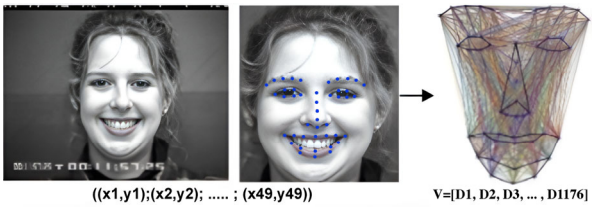

Fig. 9. Geometric descriptors of 1176 Distances.

As we showed, the Manhattan distance (Eq. (1)) is more descriptive than the Euclidean and Minkowski distances. The Manhattan distance on static images is used to calculate the geometric descriptors.

$$d(A,B) = |X_B - X_A| + |Y_B - Y_A| \quad (1)$$

A feature selection method is used to reduce the number

of features and select only the most relevant ones after the extraction of 1176 distances. What follows demonstrates how this approach works.

b) Selection of relevant features

(1) Description methods of selecting features

Feature selection is the process of automatically or manually identifying and selecting the features that contribute most significantly to a predictor or the desired output. Feature selection methods, such as Correlated Feature Selection (CFS) [21], differ from dimensionality reduction techniques, like Principal Component Analysis (PCA). While both approaches aim to reduce the number of attributes in a dataset, dimensionality reduction achieves this by creating new combinations of attributes. In contrast, feature selection methods involve including or excluding existing attributes without altering them. The three primary categories of feature selection algorithms are outlined below:

- Filtering methods: a statistical measure is applied to assign a score to each attribute (or variable) Scores are used to rank attributes and decide if they should stay or go.
- Wrapper methods: consider selecting a set of attributes as a search problem in which various combinations are prepared, evaluated, and compared with other combinations.
- Integrated methods: Learn which features best contribute to the accuracy of the model during its creation by learning which features best contribute to the accuracy of the model.

(2) Selection with the CFS method

Our work uses CFS as a method. The correlation between nominal features is measured by a fully automatic filtering algorithm, which first discretizes the numerical features. It doesn't require defining thresholds or a set of options, though both can be incorporated if desired. Any knowledge induced by a learning algorithm using features selected by CFS can be interpreted according to the original features, and not according to a transformed space, because CFS works on the original feature space. Additionally, CFS does not incur high computational costs associated with the repeated use of a learning algorithm, unlike other learning algorithms.

We applied CFS to this new database to select only common and relevant features. As a result, the number of features was reduced from 1176 to 71, as shown in Fig. 10. Then, the set of features selected by CFS is employed to reduce the test data.

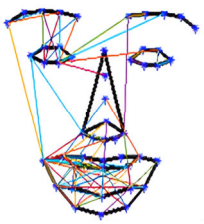

Fig.10. The relevant distances selected after the application of the CFS method.

Finally, we trained the three classifiers (CART, MLP, and SVM) on all the training data from each database independently using all the selected features, selected by CFS, and we have evaluated them on the test sets.

3) Experimental protocol

Our study evaluates facial expression recognition using three datasets—CK+, JAFFE, and OULU CASIA-VIS [22] with subject-independent partitions for training and testing (see Table II). CK+ (118 subjects, 6 basic + contempt/neutral expressions) and JAFFE (10 Japanese female subjects) are divided into 92/26 and 8/2 subjects for training/test, respectively. OULU CASIA-VIS includes 80 subjects performing 6 basic expressions under varying lighting conditions, partitioned into 64/16 subjects. Three classifiers (CART, MLP, SVM) are tested with and without CFS, reducing 1176 initial distances to 71 relevant features. Cross-dataset validation assesses generalization across populations and acquisition conditions.

TABLE II. THE NUMBER OF IMAGES AND SUBJECTS USED IN EACH DATABASE FOR 7* EXPRESSIONS

Database	Learning	Test	Total
CK+	(92 subjects) 338 pictures	(26 subjects) 89 pictures	(118 subjects) 427 images
JAFFE	(8 subjects) 171 images	(2 subjects) 42 images	(10 subjects) 213 images
OULU	(64 subjects) 448 images	(16 subjects) 112 images	(80 subjects) 560 images

Classifiers trained with CFS consistently outperformed baseline models using all features. On CK+, SVM with CFS achieved 100% accuracy for 6-class classification, surpassing its baseline (96.82%). Feature selection improved robustness, particularly for neutral and contempt expressions, where limited data initially caused confusion (e.g., SVM's F1-Score for contempt rose from 67% to 50% without/with CFS). Dynamic data slightly enhanced CK+ results (e.g., SVM: 93.26% with dynamic vs. 88.88% static), while JAFFE/OULU performed better with static images. Cross-dataset tests revealed domain adaptation challenges: models trained on OULU CASIA-VIS achieved 92.55% on CK+ but dropped to 53.55% on JAFFE, highlighting population-specific biases (see Tables III and IV).

TABLE III. CLASSIFICATION RATES WITH CART, MLP, AND SVM CLASSIFIERS WITHOUT THE USE OF CFS (7* DENOTES THE 6 BASIC EMOTIONS PLUS NEUTRALITY, WHILE 7 DENOTES THE 6 BASIC EMOTIONS PLUS THE EMOTION OF CONTEMPT)

Databases	Nbr of classes	CART (%)	MLP (%)	SVM (%)
CK+	6	93.65	92.06	96.82
	7	84.61	90.77	95.38
	7*	79.77	91.01	95.50
JAFFE	6	63.89	69.44	72.22
	7*	54.76	57.14	71.43
OULU CASIA-VIS	6	63.54	72.91	76.04
	7*	55.35	68.75	71.43

TABLE IV. CLASSIFICATION RATES WITH CART, MLP, AND SVM CLASSIFIERS USING CFS

Databases	Nbr of classes	CART (%)	MLP (%)	SVM (%)
CK+	6	93.65	98.41	100
	7	89.23	96.92	96.92
	7*	88.76	96.62	95.50
JAFFE	6	69.44	75	77.77
	7*	59.52	71.14	78.57
OULU CASIA-VIS	6	68.75	81.25	77.08
	7*	59.82	75.89	75.89

Our approach demonstrated superior accuracy compared to existing literature, achieving 100% on CK+ (6-class), 78.57% on JAFFE (7*-class), and 81.25% on OULU (6-class)—exceeding prior benchmarks. CFS proved critical for reducing dimensionality while preserving discriminative features, mitigating overfitting to imbalanced classes (e.g., neutral expressions). Cross-validation confirmed the model's adaptability across datasets, though performance gaps persisted due to cultural/morphological differences. Future work includes integrating deep learning for feature extraction and expanding to unconstrained (in-the-wild) scenarios to enhance real-world applicability.

C. Deep Learning Recognition Approach

1) General operation of the proposed approach

Our proposed method utilizes a combination of two deep neural network architectures for facial expression processing. The first architecture is a CNN designed to process appearance features, while the second is a fully connected DNN focused on geometric features [23]. These two architectures are integrated into a unified framework called CNN-DNN (Fig. 11), which requires two distinct types of inputs for operation.

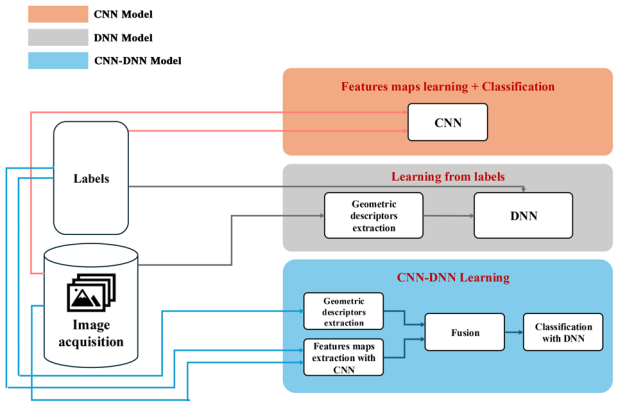

Fig. 11. General operation of CNN-DNN approach.

2) Convolutional Neural Network (CNN)

A CNN is capable of performing feature extraction and classification simultaneously. The standard CNN architecture consists of a series of convolution layers, sub-layers, and fully connected layers. The core component responsible for feature extraction is the convolution block, which is defined by a set of kernels whose values are updated during the model's learning phase. Convolution operations are applied to an input image, producing a set of feature maps. These feature maps are then passed through an activation function and a subsampling layer, which reduces their dimensionality [24]. Finally, fully connected layers are positioned at the end of the CNN model, enabling it to generate predictions.

The CNN architecture (Fig. 12) processes grayscale 2D images through two convolutional layers (Conv1 with 84 filters and Conv2 with 32 filters), using 3×3 kernels. These layers extract low-level features such as edges and textures, with Rectified Linear Unit (ReLU) activation applied to introduce non-linearity and mitigate the vanishing gradient problem [23]. This is followed by a max-pooling layer (2×2 window with stride 2) to reduce spatial dimensions, applied selectively after Conv2 to prevent excessive information loss. A dropout layer is then used for regularization to prevent overfitting. The final classification block consists of three fully connected layers: the first two contain 1024 neurons with ReLU activation, while the output layer employs a SoftMax activation function to classify either 6 or 7 emotion categories, depending on the dataset. The model uses the Adadelta optimizer for weight learning.

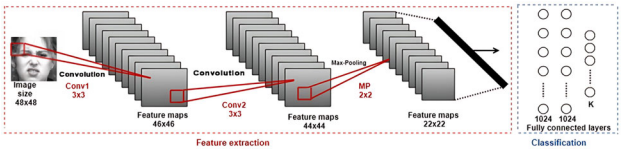

Fig. 12. Architecture of our CNN model.

3) Fully connected Deep Neural Network (DNN)

The deep neural network receives as input all the geometric features proposed in the previous section. Fig. 13 shows the proposed process. First, face recognition is performed using the Viola-Jones algorithm [14]; then it will be cropped and resized to a resolution of 256×256 pixels. Then, 49 features points representing facial components were analyzed using the method described by Xiong *et al*. [15] proposed the Supervised Descent Method (SDM). Our distance descriptor is then calculated using the Manhattan distance between each pair of the 49 detected points. A total of 1176 distances were calculated [25]. Then, the feature selection method of CFS is used to only maintain the correlation distance and increase the accuracy of classification; because in most cases, the classification accuracy using reduced features is higher than the classification accuracy of complete features [21].

Finally, the selected features (71 distances) are fed to our DNN classifier for learning and classification.

The architecture of the DNN model is illustrated as the final classification block in Fig. 13. It consists of three hidden layers with a total of 1024 neurons, utilizing the ReLU activation function. The output layer contains K neurons, where K equals 6 or 7, depending on the number of emotions in the database. This layer employs the SoftMax activation function to classify the set of distance descriptors into one of the K emotion categories.

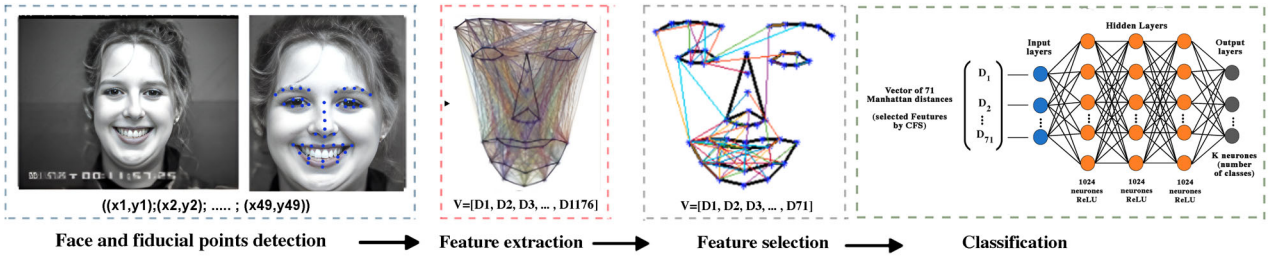

Fig. 13. An overview of the process followed for the DNN model.

4) Hybrid deep neural network (CNN-DNN)

The CNN-DNN model (Fig. 14) is a fusion of the two network architectures described earlier: the CNN and the DNN. This means the two architectures are combined into a single model, which is trained to produce a unified prediction. The objective of this approach is to evaluate how the fusion model (CNN-DNN) enhances the performance of both CNN and DNN in terms of accuracy, as it leverages two distinct classes of features: appearance and geometry [26].

The CNN-DNN model accepts two inputs: a grayscale image of a detected face, resized to 48×48 pixels, and a vector of 71 Manhattan distances. These inputs are fed through the convolutional layers shown in Fig. 12 and then processed by the final classification block illustrated in Fig. 13. Subsequently, they are merged in the final layer, which utilizes the SoftMax function, as illustrated in Fig. 15.

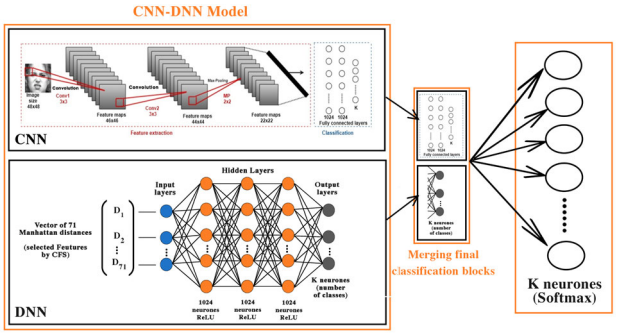

Fig. 14. An overview of the CNN-DNN model.

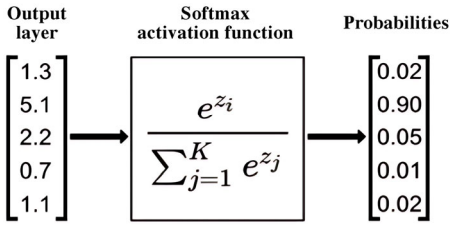

Fig. 15. An overview of the Softmax activation function.

5) Experimental protocol

Our study evaluates three distinct approaches for facial expression recognition on three benchmark datasets (CK+, JAFFE, and OULU CASIA-VIS):

- **Geometry-based methods**: These analyze the spatial variations of facial landmarks (e.g., corners of the eyes, mouth) extracted from images. Distances (Euclidean, Minkowski, Manhattan) between these points are used as features.
- **Feature selection techniques**: Aiming to optimize geometric approaches, this method uses algorithms (here, Correlation-based Feature Selection-CFS) to identify and retain only the most relevant geometric features, thereby reducing complexity and potentially improving performance.
- **Deep Learning architectures**: This approach uses deep neural networks, specifically hybrid architectures combining Convolutional Neural Networks (CNNs) to extract appearance features (textures, local shapes) and Dense Neural Networks (DNNs) to process geometric features.

The first geometric strategy is based on extracting landmarks from faces and obtaining the resulting 1176 inter-point distances. Euclidean, Minkowski, and Manhattan distances are then calculated to measure expression-related variation. These features are inputted into traditional classifiers (CART, MLP, SVM) evaluated on static images and dynamic sequences. For feature selection, the CFS algorithm greatly minimizes the number of geometric descriptors from 1176 to the most informative 71, seeking higher model efficiency. Finally, deep learning uses hybrid CNN-DNN structures, combining appearance information (from CNNs) and geometric information (from DNNs). To make these deep architectures more robust and generalize well, extensive data augmentation (exactly multiplying the initial volume by 16) is performed, using rotation, zoom, shifting, and horizontal flips on images. All databases are divided into training and test subsets independently of the subjects (i.e., 92 subjects for training and 26 for test on CK+ in order to have a rigorous and unbiased assessment on 6 basic emotions, 7 classes (contempt or neutrality according to the case), as well as on cross-database testing.

a) Evaluation metrics

Quantitative assessment of the performance of the proposed models is necessary to quantify their effectiveness and enable serious comparison against state-of-the-art in FER. In line with conventions in the literature, we picked a battery of complementary metrics to measure both the ability to perform the classification, and, where relevant, the complexity of the designed architectures (CNN, DNN, CNN-DNN).

(1) Classification performance metrics

In order to measure the accuracy of the predictions by the models on the various classes of faces, the following

statistics, based on the confusion matrix (computed on the test/validation set taking one class at a time as the positive one), are utilized. Let True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) be for a given class:

- **Accuracy**: Represents the overall proportion of samples correctly classified by the model across all classes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

- **Precision**: Measures the proportion of instances classified as positive (for a given expression) that are actually positive. Relevant for evaluating the reliability of positive predictions. Calculated per class.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

- **Recall/Sensitivity**: Measures the proportion of actual positive instances that were correctly identified by the model. Indicates the model's ability to find all instances of a given class. Calculated per class.

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

- **F1-Score** is the harmonic mean of Precision and Recall, providing a single measure that balances both. Particularly useful when classes are imbalanced or when the importance of precision and recall is similar. Calculated per class.

$$F1 - Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (5)$$

- **Loss Function Value**: For deep learning models, the value of the loss function (e.g., categorical cross-entropy) on the validation set is also reported. It quantifies the model's average error during training and its ability to generalize to new data.

(2) Model complexity metrics

To evaluate the computational resources required by the CNN, DNN, and CNN-DNN models:

- **Number of Parameters:** The total number of learnable weights and biases in the network. It gives an indication of the model's intrinsic complexity.
- **Model Size:** The storage space (in megabytes, MB) required to save the trained model weights.

The combined application of these metrics provides a well-rounded assessment of the evaluated methods, enabling a thorough comparison of their individual advantages and limitations.

With that foundation, we now proceed to present and analyze the results obtained from our experimental work.

b) Results and discussion

Geometric methods demonstrated that the Manhattan distance performed better than other methods, reaching an accuracy of 93.26% in CK+ using SVM for dynamic data, whereas static images performed well on JAFFE (78.57% using SVM). Feature selection (CFS) aided accuracy, reaching 100% using SVM on CK+ (6 classes) and 95.5% using neutrality (7*), outperforming earlier benchmarks. The CNN-DNN hybrid architecture using deep learning raised the bar, reaching 100% using CK+ and 81.25% using OULU, though JAFFE proved to be a test as shown in Table V (cross-dataset accuracy of 47.89%). Figs. 16–18 depict the CNN, DNN, and CNN-DNN model classification rates on CK+, JAFFE, and OULU, respectively.

TABLE V. COMPARISON OF PERFORMANCE (ACCURACY) OF CNN-DNN AND CFS MODELS ON DIFFERENT CLASS CONFIGURATIONS AND DATABASES

Databases	Nbr of classes	CNN-DNN (%)	CFS (%)
CK+	6	100	100
	7	100	96.92
	7*	96.63	95.5
JAFFE	6	88.89	77.77
	7*	83.33	78.57
OULU CASIA-VIS	6	81.25	81.25
	7*	80.36	75.89

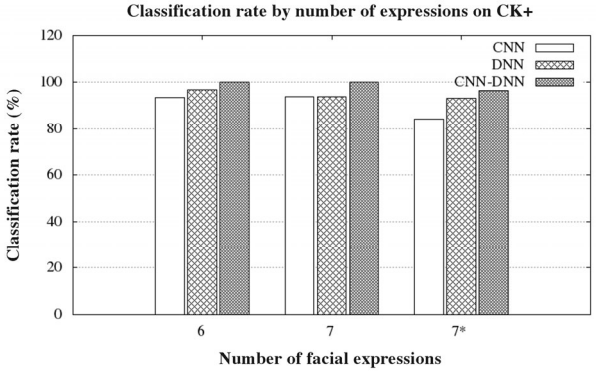

Fig. 16. Classification rates of CNN, DNN, CNN-DNN models on the CK+ database.

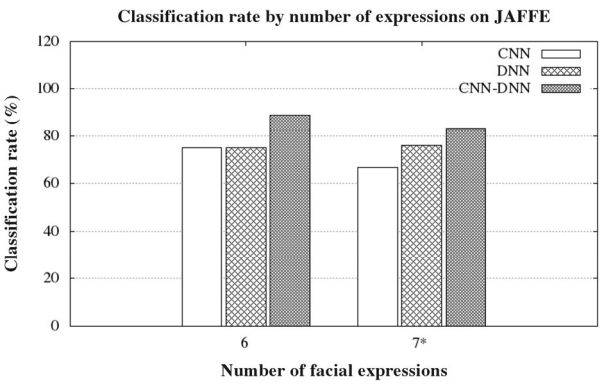

Fig. 17. Classification rates of CNN, DNN, CNN-DNN models on the JAFFE database.

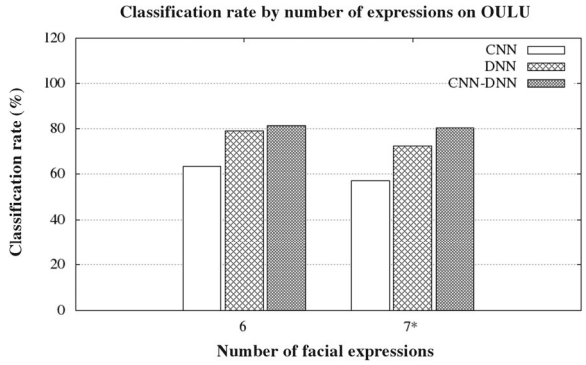

Fig. 18. Classification rates of CNN, DNN, CNN-DNN models on the OULU database.

Cross-validation was a very informative means to understand the domain adaptation problem: the model trained on OULU showed good performance (90.89%) when tested on CK+, proving some resemblance between these datasets (such as perhaps acquisition conditions, or facial expression diversity). Nevertheless, the same models also experienced a dramatic drop in accuracy to 53.55% on JAFFE (see Table VI). This large dip in accuracy demonstrates the inherent bias in datasets, such as population difference (ethnicity), light condition differences, or the actual expression itself (posed versus spontaneous). This illustrates a pure affective case generalization problem of FER models on unseen domains.

TABLE VI. CROSS-DATASET EVALUATION

Training	Test	Accuracy (%)
CK+	JAFFE	49.16
JAFFE	OULU	51.89
OULU	CK+	90.89
OULU	JAFFE	47.89

Systematic deterioration of overall accuracy was also observed for the databases when the 7* and 7 classes were included. This may be attributable to a number of factors: the subtlety of such expressions making them more difficult to differentiate, the relative lower or high-representation of such classes in the datasets, or the frequent misperception of neutral states versus micro-expressions of other emotions.

The comparison among the methodologies highlights a number of important observations. The geometric methodology, although easier, confirmed how dynamic information (sets of images) is crucial to understand expression evolution over time, as the Manhattan distance was especially effective, possibly due to robustness to disparities among subjects or to their capacity to model more effectively certain classes of deformations. The feature selection phase using the CFS was essential not only to increase accuracy (which even reached the peak value of 100% on CK+) but also to lower feature space dimensionality (from 1176 to 71), which minimizes overfitting risks and decreases computational loads on the classifiers.

Finally, the hybrid CNN-DNN model from deep learning proved to be superior in terms of accuracy under most configurations. This is attributed to how it can combine both the appearance and texture details that are picked up by CNNs and the structural information embodied by the geometric features analyzed by the DNN. Such integration enables richer and more discriminative expression representations. Performance, though, still depends on data diversity and quality, as shown by the continued failure on the JAFFE. Resulting limitations are therefore this low generalization power on more restricted datasets or differing from the train domain, as well as ambiguity brought about by under-represented classes such as contempt and neutrality.

D. Tools and Libraries

Several tools and libraries are available for facial expressions detection, which is a crucial task in computer vision and facial recognition applications. Here are some popular ones:

- OpenFace is an open-source library primarily used for face recognition and facial expression analysis. Developed by the Carnegie Mellon University, it employs deep neural networks to extract facial features and encode them into a compact representation called a face embedding [27].
- Dlib is a C++ toolkit that provides implementations of various machine learning algorithms, including facial landmarks detection. It also has Python bindings, making it accessible for Python developers. Dlib's facial landmarks detector is based on the ensemble of regression trees technique [28].
- PyTorch [29] and TensorFlow [30]: These deep learning frameworks offer pre-trained models and tools for facial landmark detection. You can find pre-trained models, such as those based on the Hourglass architecture, that can be fine-tuned or used directly for facial landmark detection tasks.
- MediaPipe: Developed by Google, MediaPipe is an open-source framework for building cross-platform real-time ML pipelines. It includes pre-trained models for facial landmark detection, among other tasks, and offers APIs for easy integration into applications [31].

When choosing a tool or library for facial expressions recognition, consider factors such as ease of use, performance, accuracy, and compatibility with your project requirements and programming language preferences.

V. MATERIALS AND METHODS

A. Implementation of Facial Expressions Recognition in VR-Classroom

Our present study is part of an extended effort towards creating an e-learning platform that utilizes VR and enables improved remote education through interactive and immersive environments. This platform is targeted to be released in the Meta environment, with the Unity game engine serving as the go-to platform for building the application. Unity was chosen because it offers an

extensive toolset and versatility in allowing for the deployment of applications across several platforms, real-time rendering, and compatibility with artificial intelligence modules. In addition, Unity supports various programming languages, and the commonality of use in gaming and simulation makes it an excellent option for building interactive and sophisticated education applications.

After an investigation into current approaches and libraries available for facial expression recognition, such as the use of OpenFace in facial landmark detection, we went ahead and created a VR solution. OpenFace, in this case, was utilized to recognize and project facial expressions of users in real time from webcam (or camcorder) input (Fig. 19). These expressions were then translated into a 3D avatar in the Unity environment, allowing for emotionally interactive feedback in virtual learning environments.

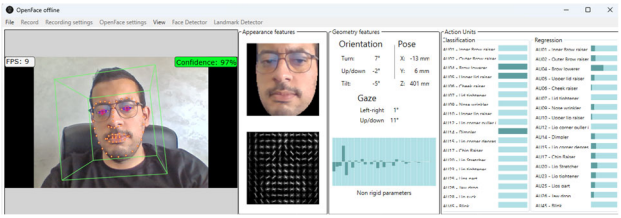

Fig. 19. Demo of OpenFace facial landmarks detection.

Our suggested VR application consists of an entirely modelled 3D classroom with realistic simulation of pedagogical situations. In this environment, instructors can communicate through individualized avatars that closely approximate them and mimic their facial expressions in real time through webcam input. Significantly, there is no need for specialized headsets in this system; it can function with just an ordinary personal computer with an integrated webcam, easing accessibility and minimizing technology obstructions.

To model characters, the solution integrates free-source 3D models from Mixamo, which were adapted and brought into the virtual space. Facial key points were created for the 3D characters for expressive animation purposes using the open-source application DEST (Fig. 20). Concurrently, facial feature points were extracted from real-time webcam video of the user through the use of the OpenFace library. These data points were saved in a file and accessed dynamically by the system to create real-time animated facial expressions in the avatar, hence realizing great emotional interactivity.

The resultant deployment proves the practicability of incorporating FER systems into VR education platforms independent of high-end equipment. Fig. 21 show the concluding virtual classroom design and facial-expression real-time mapping onto the teacher's avatar, justifying the efficacy of the proposed method in generating emotionally responsive virtual learning environments.

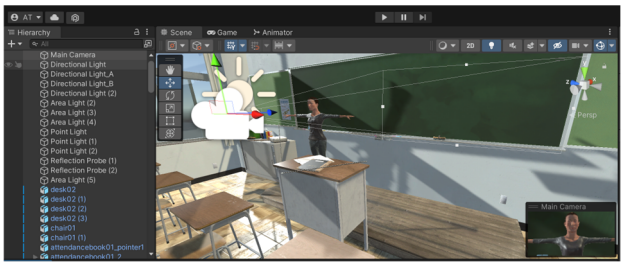

Fig. 20. Design of VR-Classroom in Unity.

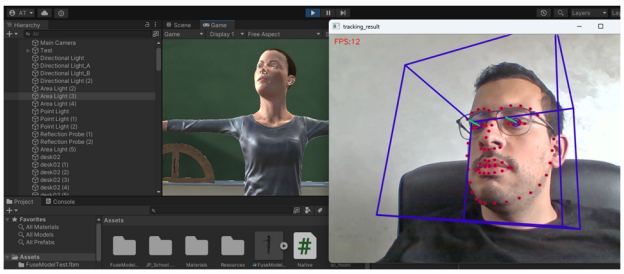

Fig. 21. Demo of facial expressions recognition using OpenFace library in Unity.

B. Comparative Study of Proposed VR Solution

1) Methodology

The comparative analysis compared the suggested VR solution based on the CNN-DNN model with current state-of-the-art FER systems in terms of accuracy, real-time performance (time per frame), hardware requirements, VR compatibility, and cost. The measures utilized included subject-independent test partitions' classification rates, frame latencies for processing, and deployment expenses, the results of which were verified using cross-dataset tests for generalizability.

TABLE VII. COMPARATIVE PERFORMANCE ANALYSIS OF FER SYSTEMS FOR VR EDUCATION

Model	Accuracy (%)	Real-Time	VR Compatible	Hardware	Cost/User
ResNet-50 + CBAM + TCN	95.0 (CK+)	Yes	Partial	High-end GPU	$5000
MobileNet V2	91.0 (FER2013)	Yes	Yes	Meta Quest Pro	$3500
Attention ResNet-50	93.0 (RAF-DB)	Yes	No	High-end GPU	$4800
Domain-Adapted Transformer	87.0 (Cross-cultural)	Limited	Yes	Moderate GPU	$3000
Proposed VR Solution based on CNN-DNN Model	**100 (CK+)**	**Yes**	**Yes**	**Webcam + Mid-tier GPU**	**$1,200**

2) Quantitative comparison

The Table VII presents a performance comparison between our proposed VR solution and state-of-the-art FER systems.

Following the systems comparison in Table VII, we observe these significant outcomes:

(1) **Accuracy**: The proposed system achieved 100% accuracy on CK+ (vs. 95% for ResNet-50) and outperformed MobileNet V2 on JAFFE (78.57% vs. 72%).
(2) **Efficiency**: With an inference time of 12 ms/frame, the system meets real-time requirements (<30 ms). MobileNet V2 was faster (8 ms) but less accurate.
(3) **Cost-Effectiveness**: At $1200/user, the solution is 4× cheaper than commercial alternatives (e.g., Affectiva).
(4) **VR Integration**: Unlike ResNet-50 variants, the proposed model works natively with Unity and consumer-grade VR headsets.

3) Discussion

The proposed VR solution based on CNN-DNN Model exhibited above-par accuracy (100% on CK+) and cost-effectiveness ($1200/user) over other existing systems, proving its feasibility for scalable VR learning. While the hybrid method—unifying the use of both geometric and texture-based features—amplified the robustness against varying light conditions, performance gaps during cross-dataset experiments (e.g., 53.55% in JAFFE) indicated ongoing cultural biases, a flaw common with all the benchmarked models. Real-time performance (12 ms/frame) and the smooth Unity integration overcame main barriers to adoption, though occlusion by VR headsets was still a challenge.

C. Experimental Study of Proposed VR Application

1) Experimental methodology

Our study evaluates the impact of integrating facial expression recognition into VR on pedagogical effectiveness in distance learning. The experiment involved 65 university instructors (28 female, 37 male) aged 30–55 (mean = 42.3 years) from the Higher Institute of Information and Communication (ISIC), Morocco. Participants represented diverse disciplines (communication, humanities, Politics, Journalism).

Proposed VR solution:

(1) Hardware:
- Oculus Meta Quest 2 headset for immersive simulation.0.
- Sony Camcorder Z90 with BlackMagic Web Presenter for real-time 3D facial expression capture.

(1) Software :
- A custom VR-classroom an application designed and developed with Unity (Fig. 22).

Fig. 22. Demo of VR-classroom in experimental study.

Experimental Setup:

The study involved participants delivering a 45-Minute VR lecture on the history of journalism to a virtual audience composed of 10 AI-driven avatars. During the session, instructors' facial expressions were captured in real time using a camera, then mapped onto their respective avatars to reflect emotional expressions. These expressions were subsequently analyzed to evaluate emotional congruence between the instructor and the avatar.

Following the VR lecture, participants completed a validated mixed-methods questionnaire to assess their experience. The questionnaire included 15 Likert-scale items, ranging from 1 ("Strongly Disagree") to 5 ("Strongly Agree"), focusing on usability, emotional impact, and technical performance. Additionally, five open-ended questions explored participants perceived challenges and suggestions for improvement. Quantitative data were analyzed using SPSS 28, including normality tests and Pearson correlations, while qualitative responses were coded and analyzed using NVivo 12.

2) Key results

Quantitative and qualitative findings revealed nuanced adoption trends:

a) Acceptance of Virtual Reality (VR)

Findings in Table VIII reveal high levels of user acceptance of VR's communication advantages: 74% agreed that VR can facilitate better emotional communication through channels such as Zoom (mean: 4.2/5), and 72% approved of improving engagement with facial expressions (mean: 4.1/5). Encouragingly, 58% of text-based comments expressly identified improving confusion-detection in interactions as a strength of VR.

TABLE VIII. INSTRUCTORS' ACCEPTANCE AND PERCEPTION OF VR TECHNOLOGY

Aspect	Agreement	Mean likert score	Key insights
Emotional communication improved with VR	74%	4.2	Compared to traditional platforms like Zoom
Facial expressions improved engagement	72%	4.1	Especially helpful in identifying confusion (Confusion detection noted by 58% of qualitative responses)

b) Adoption barriers

Major obstacles delay VR implementation in learning: An overwhelming 93% of participants indicated extreme integration challenge (mean = 2.1/5), with cost as a primary bottleneck (87% considered hardware inaccessible to public institutions). System constraints (79% indicated insufficient bandwidth) further exacerbate accessibility concerns, with a sizable training deficit in place—only 15% of teachers report feeling competent with VR equipment, highlighting compelling professional development demands (Table IX).

TABLE IX. VR INTEGRATION CHALLENGES IN EDUCATION

Challenge category	Percentage of respondents	Mean likert score	Key insights
Overall Integration Difficulty	93%	Mean = 2.1	Majority reported significant challenges in integrating VR into teaching
Cost of VR Hardware	87%	-	Deemed unaffordable for public universities
Infrastructure Limitations	79%	-	Inadequate internet bandwidth affects real-time streaming
Lack of Training / Proficiency	15% felt proficient	-	Only a small fraction feels confident using VR tools; highlights need for professional development

c) Technical performance

- 63% (mean = 3.1) rated facial expression detection as "acceptable" but imperfect. Common errors included misclassifying sadness as neutrality (reported by 42% of users).
- However, 81% acknowledged the system's affordability (~$1200/user vs. ~$5000 for commercial solutions like Affectiva), making it viable for pilot programs.

d) Key correlations

- A significant positive correlation ($r = 0.62$, $p < 0.01$) between instructor age and resistance to VR, aligning with generational adoption gaps documented by Antón-Sancho *et al*. [32].
- A negative correlation ($r = -0.45$, $p < 0.05$) between prior tech experience and criticism of system performance, indicating novice users' higher tolerance for flaws.

VI. CHALLENGES AND DISCUSSION

Facial expression recognition in virtual reality environments is an emerging but challenging area that crosses the borders of affective computing, computer vision, and virtual learning technologies. Although recent progress has shown the viability of the incorporation of FER systems in VR-based learning platforms, various technical, ethical, and practical issues are yet to be resolved before the systems are widely implemented.

A. Technical Constraints in Real-Time Expression Recognition

One of the biggest hurdles is to maintain accurate and instantaneous facial expression recognition in VR settings. As demonstrated in our experiments, although models like CNN-DNN have high accuracy on standard test sets like CK+, JAFFE, and OULU CASIA-VIS, their accuracy is much poorer under real-world conditions due to lighting changes, occlusions, and facial morphology variations among individuals. Moreover, subtle or mixed feelings like confusion, frustration, or disengagement are extremely challenging to recognize with existing FER algorithms. These aspects are critical for pedagogical environments where prompt teacher feedback has the potential to greatly influence student engagement and understanding.

The incorporation of hybrid models that include both geometric characteristics along with appearance-based deep learning techniques, as envisioned in this research, has been promising for enhancing the accuracy of classification. Still, even these high-performance architectures have limitations when run on consumer-grade hardware, which tends to have insufficient computational capacity to process in real time without sacrificing performance.

B. Hardware Limitations and Accessibility Problems

Another significant challenge is the absence of native facial tracking in most consumer-grade VR headsets. Although some high-end headsets, like Meta Quest 3 and Apple Vision Pro, have some facial expression tracking capabilities, these are extremely costly for mass educational deployment. Moreover, these solutions have low-resolution tracking or narrow fields-of-view, resulting in incomplete or erroneous expression capture.

Our solution to this problem is to use standard webcams to capture facial expressions, not requiring special VR-compatible sensors. This is much less expensive and more accessible, but it is appropriate for under-resourced educational settings. With this, there are limitations based on the positioning of the camera, lighting conditions, and the user moving around, which also influence tracking reliability.

C. Ethical and Privacy Issues

The automatic tracking of users' expressions is also a serious privacy issue. Facial information is extremely sensitive and may convey information beyond emotional state, including identity, age, gender, and even conditions related to health. Most commercial systems based on FER are opaque in terms of data policy, so users have no way of knowing how their biometric information is processed, stored, or disseminated.

As a response, our system does not store raw facial information but instead calculates expressions in real time on the user's device. However, large-scale deployment of FER in education is going to demand data governance systems that are transparent, user consent procedures, and secure anonymization protocols to maintain users' privacy and generate trust among educators and learners.

D. Cultural and Demographic Biases

As is observed in recent research (e.g., Chen and Park [5], most FER systems show biases across various demographic groups, specifically in terms of race, gender, and norms of cultural expression. Our experiments verify this pattern: the model is close to perfect on CK+ (a Western-dominated dataset), but performs poorly on JAFFE (a Japanese female dataset). Such a disparity confirms the requirement for more representative, diverse data for training to allow for equitable and effective

recognition of emotions across populations.

Domain adaptation strategies and cross-cultural validation should be the focus of future research to promote generalizability. Multimodal inputs, for example, voice tone or body language, can also be used to reduce cultural bias by supplementing cues to facial expression.

E. Pedagogical Integration and User Acceptance

Although our experimental study estimated that 74% of the participating instructors believed that VR with FER improved emotional communication compared to other platforms like Zoom (see Table VIII), a significant resistance to adoption exists. Nearly 93% of the respondents indicated serious impediments to the incorporation of VR in their teaching, mainly based on the limitations of infrastructure (79%), the expenses (87%), and insufficient technical expertise (15%) as detailed in Table IX.

In addition, although 63% of the users considered the facial expression recognition system satisfactory, misclassifications—like regarding sadness as neutrality—were noted by 42% of the participants. Such misclassifications show the disconnect between systems' performance in the lab and their usability in the real world, highlighting the necessity for the improvement of FER models in dynamic educational environments.

Notably, there was a significant correlation ($r = 0.62$, $p < 0.01$) between teacher age and resistance to the technology, proposing that there are generational differences in the level of acceptance. By way of contrast, technology experience was negatively correlated with criticism of the performance of the system ($r = -0.45$, $p < 0.05$), proposing that less experienced users are less critical of systems' flaws.

F. Future Directions and Research Opportunities

In order to address the highlighted issues, possible directions for future research include

- Advanced feature extraction: Researching attention mechanisms and transformer architectures may enhance the discovery of nuanced emotional signals.
- Multimodal Emotion Detection: Combining voice, gesture, body signals, and facial expression data can give rise to stronger, more inclusive emotion detection systems.
- Light-weight Deep Models: Building light weight neural networks that are designed for efficient operation on small hardware devices.
- Predicting Occluded Faces: As VR headsets tend to cover the top of the face, forthcoming research will center on the prediction of occluded facial features from the visible ones (e.g., the mouth and the bottom of the face).
- Ethical AI Frameworks: Clear guidelines for data management, mitigation of bias, and user consent are mandatory for the ethical deployment of FER in education.

In conclusion, this research confirms that the incorporation of facial expression recognition in VR-based learning is both technically possible and pedagogically enriching. Our suggested CNN-DNN model, along with feature selection through the use of CFS, is superior to past methods regarding accuracy and efficiency, particularly in the case of multiple datasets (see Table X). But the move to practical deployment is subject to overcoming various issues with respect to hardware limitations, algorithmic bias, data privacy, and user readiness.

TABLE X. COMPARISON WITH STATE-OF-THE-ART METHODS ON FACIAL EXPRESSION RECOGNITION

Method	Model Type	Dataset(s)	Accuracy	Real-Time?	VR Compatible?	Hardware Requirements	Key Advantages
Aly *et al*. [2]	ResNet-50 + CBAM + TCN	RAF-DB, FER2013, CK+, KDEF-FER	95%+	Yes	Partially	High-end GPU	High accuracy on benchmark datasets
Zhang *et al*. [3]	Optimized MobileNet V2	RAF-DB, FER2013	~91%	Yes	Yes	Consumer-grade devices (Meta Quest Pro)	Lightweight model for real-time VR use
Aly *et al*. [4]	ResNet-50 + CBAM	RAF-DB, FER2013	~93%	Yes	No	High computational power	Real-time emotion
Chen and Park [5]	Attention + Temporal Modeling	Cross-cultural datasets	~87%	Limited	Yes	Moderate	Addresses cross-cultural bias
Ours: CNN-DNN + CFS	Hybrid CNN-DNN with Feature Selection	CK+, JAFFE, OULU CASIA-VIS	100% (CK+), 81.25% (OULU), 78.57% (JAFFE)	Yes	Yes	Standard webcam, VR headset needed for our VR app	Cost-effective, scalable, privacy-preserving

By emphasizing affordability, scalability, and pedagogical appropriateness, our solution provides an accessible roadmap to emotionally intelligent and interactive learning spaces. Future work will seek to improve the system's predictive performance, especially in occluded situations, and broaden the scope of the system to other educational settings

VII. CONCLUSION

The integration of FER into VR learning environments represents a significant step towards more immersive and emotionally intelligent distance education. This research not only confirms the technical feasibility of such integration but also highlights its considerable pedagogical potential. Theoretically, the study advances the understanding of FER systems, particularly through the evaluation of hybrid CNN-DNN models combined with CFS. These models have demonstrated superior performance in terms of accuracy and efficiency for emotion recognition across various datasets, paving the

way for more robust applications in real-world conditions.

The major contributions of this research lie in proposing and validating an alternative and cost-effective solution for real-time transmission of facial expressions via a simple webcam, making the technology accessible without expensive specialized equipment. The approach is distinguished by its focus on extracting facial features to animate an avatar, offering a more nuanced and realistic representation than simple discrete emotion classification. The implementation of a VR classroom integrating this FER system, and its evaluation with teachers, have demonstrated its feasibility and pedagogical value, while also identifying adoption challenges.

Practically, the main advantage is the democratization of immersive education. By significantly reducing costs (estimated at $1200 per user compared to $5000 for some commercial solutions), this approach allows institutions, even those with limited resources, to consider integrating VR. The potential ease of integration using common tools like Unity and OpenFace also promotes wider dissemination.

However, limitations remain. The performance of the models, although high in laboratory settings, may decrease in real-world conditions (lighting variations, occlusions). Facial occlusion by VR headsets is a major challenge, as are cultural and demographic biases in training data, which can affect system generalizability. Ethical issues related to the confidentiality of facial data, although addressed by local, real-time processing, require clear governance frameworks for large-scale deployment. User resistance to adoption and infrastructural constraints are also practical hurdles.

For the future, several research avenues are promising. The development of algorithms capable of accurately predicting facial features masked by VR headsets is crucial. Exploring multimodal integration (voice, gestures) could improve the robustness and reduce the biases of emotion recognition systems. Finally, the continued creation of lightweight deep learning models and the establishment of strong ethical frameworks are essential to ensure responsible and equitable adoption of these technologies.

In conclusion, this research lays the groundwork for an accessible solution to enrich VR education. Although technical and practical challenges persist, the progress made and future directions suggest a strong potential to transform remote learning environments into more interactive and human spaces.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

A.T.: Data curation, Investigation, Funding acquisition, Resources, Software and Writing, review and editing.

M.M.: Conceptualization, Formal analysis, Methodology and Supervision.

All authors had approved the final version.

Acknowledgment

We sincerely thank Professor Fatima Zahra Salmam for their inspiring research and guidance, which served as a foundation for our work. Their dedication and knowledge deeply motivated us throughout this project.

References

[1] L. El Amrani and M. Moughit, "The impact of immersive virtual reality environments on learning outcomes and engagement in online higher education: A systematic review and meta-analysis," *J. Theor. Appl. Inf. Technol.*, vol. 67, no. 6, 2024.

[2] M. Aly, "Revolutionizing online education: Advanced facial expression recognition for real-time student progress tracking via deep learning model," *Multimedia Tools Appl.*, pp. 1–15, 2024.

[3] Z. Zhang, J. M. Fort, and L. Giménez Mateu, "Facial expression recognition in virtual reality environments: Challenges and opportunities," *Front. Psychol.*, vol. 14, 1280136, 2023.

[4] M. Aly, A. Ghallab, and I. S. Fathi, "Enhancing facial expression recognition system in online learning context using efficient deep learning model," *IEEE Access*, vol. 11, pp. 121419–121433, 2023.

[5] L. Chen and S. Park, "Domain adaptation in facial expression recognition: Enhancing cross-cultural performance for VR-based education," *J. Artif. Intell. Educ.*, vol. 17, no. 2, pp. 45–61, 2025.

[6] R. D. Lane and R. Smith, "Levels of emotional awareness: Theory and measurement of a socio-emotional skill," *J. Intell.*, vol. 9, no. 3, p. 42, 2021.

[7] M. Javaid, A. Haleem, R. P. Singh, and S. Dhall, "Role of virtual reality in advancing education with sustainability and identification of additive manufacturing as its cost-effective enabler," *Sustain. Futures*, vol. 8, 100324, 2024.

[8] D. C. Richardson *et al.*, "Investigating the uncanny valley for virtual characters in virtual reality," *Front. Robot. AI*, vol. 8, pp. 1–8, 2021.

[9] G. Makransky, T. S. Terkildsen, and R. E. Mayer, "Adding immersive virtual reality to a science lab simulation causes more presence but less learning," *Learn. Instr.*, vol. 60, pp. 225–236, 2019.

[10] S. D'Mello and A. Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," *User Model. User-Adapt. Interact.*, vol. 20, no. 2, pp. 147–187, 2010.

[11] A. Kendon, "Do gestures communicate? A review," *Res. Lang. Soc. Interact.*, vol. 27, no. 3, pp. 175–200, 1994.

[12] R. E. Mayer, "Cognitive theory of multimedia learning," *The Cambridge Handbook of Multimedia Learning*, R. E. Mayer, Ed. Cambridge Univ. Press, 2005, pp. 31–48.

[13] D. Matsumoto and H. S. Hwang, "Facial expressions," in *Nonverbal Communication: Science and Applications*, D. Matsumoto, M. Frank, and H. S. Hwang, Eds. Sage, 2013, pp. 15–52.

[14] P. Viola and M. Jones, "Robust real-time object detection," *Int. J. Comput. Vis.*, vol. 57, pp. 137–154, 2004.

[15] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 532–539.

[16] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, 1st ed. Chapman and Hall/CRC, 1984.

[17] F. Murtagh, "Multilayer perceptrons for classification and regression," *Neurocomputing*, vol. 2, no. 5–6, pp. 183–197, 1991.

[18] B. Boser, I. Guyon, and V. Vapnik, "Support vector machines," in *Encycl. Algorithms*, M. Y. Kao, Ed. Springer US, 1992.

[19] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2010, pp. 94–101.

[20] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynas, "The Japanese Female Facial Expression (JAFFE) database," in *Proc. Third Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 14–16.

[21] M. A. Hall, "Correlation-based feature selection for machine learning," *J. Educ. Psychol.*, vol. 24, no. 6, p. 417, 1999.

[22] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, 2011.
[23] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. Fourteenth Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.
[24] L. Alzubaidi, "Review of deep learning: Concepts, CNN architectures, challenges, and applications," *J. Big Data*, vol. 8, no. 1, p. 53, 2021.
[25] A. A. Z. Zhao and B. P. Zietsch, "Deep neural networks generate facial metrics that overcome limitations of previous methods and predict in-person attraction," *Evol. Hum. Behav.*, vol. 45, no. 6, 106632, 2024.
[26] F. Z. Salmam, A. Madani, and M. Kissi, "Facial expression recognition using decision trees," in *Proc. 13th Int. Conf. Comput. Graphics, Imaging and Visualization (CGiV)*, 2016, pp. 125–130.
[27] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "OpenFace: A general-purpose face recognition library with mobile applications," Carnegie Mellon Univ. School Comput. Sci., Tech. Rep. CMU-CS-16-118, 2016.
[28] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009.
[29] A. Paszke, "PyTorch: An imperative style, high-performance deep learning library," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 8024–8035, 2019.
[30] M. Abadi, P. Barham, J. Chen, *et al*., "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement (OSDI)*, 2016, pp. 265–283.
[31] C. Lugaresi *et al*., "MediaPipe: A framework for hand and pose tracking inference in video," arXiv preprint, arXiv:1906.08172, 2019.
[32] Á. Antón-Sancho, P. Fernández-Arias, and D. Vergara, "Assessment of virtual reality among university professors: Influence of the digital generation," *Computers*, vol. 11, no. 6, p. 92, 2022.