Hybrid CNN-DNN Facial Expression Recognition in an Affective Virtual Reality (VR) and 360° E-learning Platform

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

[1] Laboratory Science and Technology for Engineering (LaSTI), National School of Applied Sciences, Sultan Moulay Slimane University, Khouribga, 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.); m.moughit@usms.ma (M.M.)
*Corresponding author

***Abstract*—Immersive Virtual Reality (VR) is increasingly used to support experiential learning and social presence in higher education, while 360° video offers lightweight immersive experiences accessible from standard browsers and mobile devices. However, few platforms integrate deep learning-based Facial Expression Recognition (FER) with VR and 360° e-learning in a unified architecture tailored for distance education. This paper presents an integrated e-learning platform that combines: a Unity-based multi-user VR classroom, a web-accessible 360° video learning module, and a hybrid Convolution Neural Network-Deep Neural Network (CNN-DNN) FER service that fuses appearance cues from grayscale face images with geometric descriptors from facial landmarks. We describe a proof-of-concept implementation and evaluate the FER component on public datasets (Cohn-Kanade dataset (CK+), Japanese Female Facial Expression (JAFFE), University of Oulu-Institute of Automation, Chinese Academy of Sciences (OULU-CASIA)) using subject-independent splits, and we report an exploratory pilot study with 65 university instructors who experienced a short VR lecture scenario with FER-driven avatar facial animation and provided usability and perception feedback. The fused CNN-DNN achieved 100% accuracy on CK+ (6- and 7-class settings), 88.89% on JAFFE (6 classes), and 81.25% on OULU-CASIA (6 classes), while cross-dataset performance dropped to 47.89% in the OULU→JAFFE setting, highlighting remaining generalization challenges. In the pilot study, 74% of instructors agreed that the VR scenario supported emotional communication, and 72% reported that avatar facial cues helped them notice confusion or engagement, although 93% also identified substantial integration barriers related to cost, infrastructure, and training. These results support the technical feasibility of the platform while clarifying the practical and ethical conditions required for responsible affect-aware immersive learning.**

***Keywords*—virtual reality, 360° video, e-learning, facial expression recognition, deep learning, Convolution Neural Network-Deep Neural Network (CNN-DNN), affective computing**

I. Introduction

"Being in the moment" and the contextualization of learning activities have become more pressing in the wake of online and blended learning due to the emergence of social presence and emotional engagement as core factors in this matter [1]. Immersive Virtual Reality (VR) has the potential to create "being there" experiences that are difficult to achieve in the current learning management systems for learning and training activities [2]. Simultaneously, 360° video content can be experienced in a light immersion approach in any browser or in head-mounted displays that are cost-effective and can be adopted in any institution for immersive learning [3].

Recent meta-analyses and systematic reviews (2024–2025) report generally positive learning outcomes for immersive VR in higher education contexts [4]. Comparable evidence in professional training (e.g., VR simulation for robotic surgery) also indicates benefits for training but stresses heterogeneous study designs and the need for evidence-based pedagogical alignment and standardized evaluation in authentic deployments [5]. Parallel work on learning analytics for immersive virtual learning environments highlights that spatial interaction traces, gaze, and affect-related signals can support scalable feedback and adaptation but require principled multimodal analytics and careful reporting practices [6].

In parallel, Facial Expression Recognition (FER) has matured considerably with the adoption of deep learning, particularly convolutional and hybrid Convolution Neural Network-Deep Neural Network (CNN-DNN) architectures [7]. For instance, deep facial emotion recognition systems possess the capability to recognize learners' affective states such as confusion, boredom, and engagement through the webcam interface. The use of appearance-based and geometric features for FER systems has shown enhanced robustness on benchmark databases [8]. However, the majority of FER-based learning platforms are limited to 2D video conferencing

Manuscript received December 23, 2025; revised January 27, 2026; accepted March 26, 2026; published July 21, 2026.

or require sophisticated hardware, which restricts the integration of such systems with affordable virtual reality learning environments [9].

Within FER, recent surveys synthesize current trends and outline remaining gaps for real-time and education-facing systems, including interpretability and dataset considerations for responsible deployment [7]. In parallel, "in-the-wild" FER methods increasingly target robustness to pose, illumination, and occlusion and leverage attention mechanisms and transformer-style backbones for improved generalization [10]. A recent review specifically examines facial expressions in VR-based education, detailing recognition approaches and integration considerations for immersive learning experiences [11].

In our earlier prototype, we introduced a low-cost multi-user VR classroom where a webcam-based CNN-DNN FER pipeline drove avatar facial animation in real time [11]. The present paper extends that prototype by integrating a 360° web-based learning module within the same architecture, expanding the experimental evaluation and reporting, and providing a more structured user-study analysis focused on social/emotional presence and deployment constraints.

We present a deep learning-based FER module integrated into a multi-modal immersive e-learning platform that combines a synchronous VR classroom with avatars and FER-driven expressions, and a 360° video e-learning module accessible via web browser or VR headset.

The main contributions of this work are fourfold. First, we propose a unified platform architecture that combines a multi-user VR classroom, a browser-based 360° video learning module, and an FER service designed for deployment on commodity institutional hardware (personal computers, webcams, and affordable headsets). Second, we detail a hybrid CNN-DNN FER pipeline that fuses facial-landmark geometry with grayscale appearance cues and can operate in real time in a Unity-based environment. Third, we describe the integration of FER outputs into both the VR classroom (for avatar facial animation) and the 360° learning module (for time-stamped affective event logging). Finally, we provide a preliminary evaluation on public FER datasets and an exploratory pilot user study with university instructors to assess feasibility, usability, and perceived pedagogical value.

In light of these objectives, this study addresses the following Research Questions (RQ):

- RQ1: To what extent does the proposed CNN-DNN facial expression recognition model improve recognition accuracy compared to geometric baselines on benchmark datasets?
- RQ2: How do university instructors perceive the added value and challenges of a VR classroom enriched with FER-driven avatar animation and analytics?
- RQ3: Under what constraints can the integrated VR/360° platform be deployed in resource-limited higher-education contexts?

The remainder of this paper is organized as follows: Section II reviews related work on VR/360° e-learning and deep FER. Section III describes the materials and methods, which include the architectural design of the platform, the proposed CNN-DNN-based FER technique, its implementation into the VR classroom and 360° module environment, as well as the experimental procedure details and the adopted testing approach. The results and findings of the pilot study are given in Section IV, followed by Section V discussion of the findings, limitations, and the ethical issues of the study. The conclusion is offered in Section VI.

II. Literature Review

A. Immersive VR and 360° Video in Education

Systematic reviews have shown that immersion in Virtual Reality (VR) can be utilized for conceptual understanding, skill acquisition, and motivation within STEM and professional education when aligned with learning outcomes [2]. Factors that have been shown to play an important role in learning and immersion in virtual environments are embodied interaction, spatialized feedback, and narrative frameworks. Recent higher-education focused syntheses report accelerated adoption of Extended Reality (XR) across disciplines but identify persistent gaps in experimental rigor, assessment alignment, and reporting transparency [12]. In teacher education, meta-analytic evidence indicates generally positive yet heterogeneous effects of VR-based interventions, reinforcing the need for contextualized implementation and fidelity reporting [13].

More recent evidence in immersive learning specifically examines procedural training with head-mounted displays (including 360° video representations) and reports a positive overall effect while noting methodological variability and limited between-subject evidence, reinforcing the importance of clearly specified evaluation protocols [3]. Meta-analytic results focusing specifically on 360° video and spherical video VR suggest that learning gains are not always reliable across contexts and depend strongly on task type and study design [14]. A complementary meta-analysis across 360° video studies reports moderate average effects on learning and non-cognitive outcomes but also substantial heterogeneity, underscoring the need for careful instructional integration and reporting [15].

Recent reviews summarize VR as an educational tool, highlighting both its potential and the practical challenges of integrating immersion into effective pedagogy [16]. Empirical work further suggests that immersive VR can support engagement and learning when instructional design manages cognitive load and follows established multimedia principles [17]. Beyond general VR, systematic evidence on gamified VR learning environments reports consistent benefits for motivation and engagement when game mechanics are aligned with instructional objectives [18]. In language learning, XR interventions have also been reviewed as supporting contextualized practice and learner engagement, though

outcome measures and comparators vary widely across studies [19].

360° video finds itself in an in-between role in the reality-virtual continuum. Monocular or Stereoscopic panos can be navigated in a 360° video in the same way that head-tracking in head-mounted displays allows navigation of virtual environments. 360° video applications include virtual school field trips, medical walks, and training sessions for workers, offering context and lower costs of development compared to fully interactive applications in VR environments [6]. Within the realm of education, it is common to use 360° video within Internet-based players.

Comparative analysis of VR and videoconferencing has indicated that VR can potentially improve social presence, feelings of community, and perceived course learning, though this may be tempered when used for extended periods of time due to fatigue and logistical considerations [20]. For this reason, hybrid methods of incorporating shorter VR or 360-degree views within more comprehensive online learning systems and workflows may be indicated.

B. Deep Learning-Based Facial Expression Recognition

FER has made rapid progress in deep learning, moving away from hand-engineered approaches and towards end-to-end CNN-based and hybrid approaches [8]. The main difficulties in this area are intra-class variation, occlusion, head pose variation, and bias in the training data. Review of modern approaches reveals that many different architectures and training methods are being used to counter such issues [21].

Hybrid approaches combine appearance-based cues from grayscale or colour images with geometry-based cues derived from facial landmarks, improving robustness on controlled benchmarks such as Cohn-Kanade dataset (CK+) and Japanese Female Facial Expression (JAFFE) [22]. These designs aim to balance accuracy and computational cost for real-time deployment.

In parallel, compact CNN architectures such as the MobileNet family have been employed for resource-constrained or real-time FER scenarios, including student engagement monitoring for online education settings [23]. The use of FER for e-learning settings demonstrates the potential of the approach for monitoring student engagement and affect while watching live or pre-recorded lectures.

C. Affective Computing in Immersive Learning

Affective computing aims to recognize and respond to learners' affect using multimodal cues (e.g., facial expressions, voice, physiology, and interaction patterns). In immersive learning, affect-aware feedback can support adaptive scaffolding, but deployment in VR raises practical issues such as partial face occlusion from HMDs and privacy-aware data handling [24]. Experimental studies on facial affect recognition in immersive VR also highlight that gaze direction and visibility of key facial regions can affect recognition reliability, emphasizing sensor placement and interaction design [25]. Broader multimodal emotion recognition surveys further suggest that combining facial cues with voice and physiological signals can improve robustness when the face is partially occluded or when privacy constraints limit video capture [26].

In online and blended education, a review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (2019–2024) reports that facial expressions remain the dominant modality for affect detection, yet classroom validation and ethical reporting are often limited—motivating more rigorous, privacy-aware evaluations in educational settings [27]. Complementary reviews of affective computing for learning similarly emphasize multimodal sensing, responsible data handling, and cross-context generalization as key directions [28].

Our work follows this line by deploying a webcam-based FER module as a central service that feeds both a VR classroom (with or without headsets) and a 360° video learner interface, while explicitly targeting accessible hardware and open datasets.

D. Privacy-Preserving, Fair, and Multi-Modal Affect Sensing

Finally, commercial affect-sensing solutions provide packaged SDKs for facial action units and emotion estimates (e.g., Affectiva's AFFDEX/FACET toolkits), though their models and training data are typically proprietary [29]. FaceReader is another widely used commercial AFEA system and has been evaluated in validation studies and methodological reviews.

A second, equally important line of work concerns bias and fairness in FER. Methodological studies propose metrics and taxonomies to quantify demographic bias in FER datasets and to guide subgroup-aware evaluation and dataset selection [30]. More broadly, prior work on facial analysis has documented demographic performance disparities (e.g., across gender and skin tone) and motivated routine subgroup reporting and fairness checks [31]. Building on this, recent FER-specific studies distinguish representational versus stereotypical gender bias and examine how each type of bias propagates from datasets into model predictions [32]. To reduce intrusiveness, learning analytics systems can also incorporate complementary modalities such as voice prosody and interaction logs. Our architecture is modular and can integrate such multimodal cues, enabling less-invasive configurations and more robust evaluation across diverse learner populations.

Privacy-preserving FER has therefore become an active research direction. Perturbation-based approaches can obfuscate sensitive affective attributes while preserving downstream utility, including universal feature-space perturbations designed to generalize across recognition models [33]. On-device and edge implementations further reduce data exposure by keeping inference local to the learner device, enabling real-time FER without transmitting raw facial data [34]. Federated learning can additionally support personalization for video-based FER while retaining user data on-device, improving performance via collaborative training without centralized data collection [35].

Overall, the literature reviewed in this paper points to four gaps that this paper addresses. First, existing works have mostly addressed either virtual reality or 360° video in isolation, but not within a comprehensive educational platform. Second, research in affective computing and FER has continued to emphasize benchmarks over deployable and educational-focused integration within commodity hardware. Third, accessible affect-aware learning environments have seldom integrated real-time avatar expressiveness with session-level analytics within a single architectural framework. Fourth, concerns around privacy, fairness, and responsible use have often been emphasized but not necessarily translated into actual deployment within educational environments. In this paper, we propose a comprehensive educational platform that integrates a deployable and affordable VR/360° video platform with a modular CNN-DNN FER service, along with a cautious deployment framework.

III. Materials and Methods

A. Methodological Roadmap

We first describe the platform architecture and data flow (Fig. 1), then the CNN-DNN Facial Expression Recognition (FER) pipeline (Figs. 2–6). Next, we explain how FER predictions are consumed by the Unity VR classroom and the 360° web module (Fig. 7–12). Finally, we outline the evaluation protocol: offline benchmarking on public datasets and a VR pilot study.

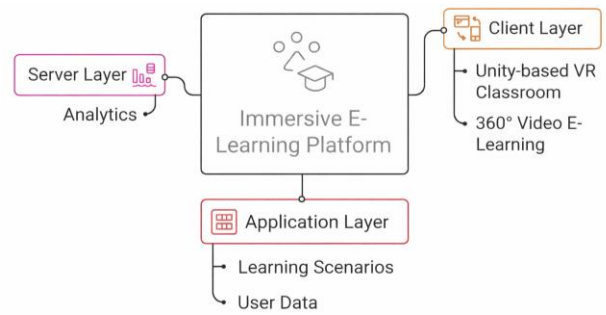

Fig. 1. Immersive VR e-learning architecture.

B. System Overview

The overall architecture of the proposed immersive e-learning platform includes three main layers (Fig. 1). At the client layer, learners can access the system through two complementary modes. A Unity-based VR classroom client enables multi-user 3D interaction with avatars, voice chat, and FER-driven facial animation when using a VR headset. In parallel, a 360° video e-learning module is delivered through a web-based player that can be experienced either with a basic 3D viewing headset (e.g., smartphone-based viewer) or directly in a standard web browser via WebXR or a native player. The upper application and server layers manage the learning scenarios, user data, and analytics.

At the application layer, a standalone FER service exposes a CNN-DNN model via REST or WebSocket, receiving cropped faces or landmark sets and returning discrete emotion labels and optional affective indicators. The session manager handles user authentication and room assignment. The analytics engine processes FER events, interaction logs, and task performance.

In the data layer, an affective events store stores time-stamped predictions of emotions regarding particular learning activities. A content repository stores VR scenes, 3D models, 360-degree videos, and quizzes. A user database stores user information, user consent, and anonymized user identifiers.

C. Deep Learning FER Pipeline (CNN-DNN)

1) General operation of the proposed approach

The proposed FER pipeline uses a dual-branch deep model, where two different face representations are learned simultaneously. The CNN branch uses appearance features extracted from grayscale face images, while the fully connected branch (DNN) uses geometric features extracted from facial landmarks [22]. The output of both branches is combined to make a final decision on the emotion, and the proposed CNN-DNN hybrid framework is used throughout the work (Fig. 2).

2) Convolutional Neural Network (CNN)

CNNs are popular choices for vision tasks owing to their ability to learn hierarchical image representations directly from image data. In a CNN, convolutional kernels are typically slid over the image to compute feature maps, which are further passed through nonlinear transformations and pooling operations to reduce spatial resolution while preserving discriminative features. Eventually, one or more fully connected layers are used to compute class scores from the learned representation.

As illustrated in Fig. 3, our CNN takes a grayscale face image and applies two convolutional blocks (Conv1: 84 filters; Conv2: 32 filters) with 3 × 3 kernels to capture local texture and edge information. Rectified Linear Unit (ReLU) activations are used to introduce non-linearity and to stabilize optimization by reducing vanishing-gradient effects. A 2 × 2 max-pooling operation (stride 2) is applied after Conv2 to reduce spatial dimensionality without overly discarding discriminative details. We then apply dropout for regularization. The classification head contains two dense layers of 1024 units (ReLU) followed by a SoftMax output layer that predicts either 6 or 7 emotion classes, depending on the dataset. Network weights are trained with the Adadelta optimizer. Implementation details added for reproducibility: we use valid 3 × 3 convolutions on 48 × 48 grayscale inputs, yielding 46 × 46 × 84 and 44 × 44 × 32 feature maps, followed by 2 × 2 max-pooling (stride 2) to 22 × 22 × 32; dropout is applied before the dense layers. Training uses Adadelta with default settings (learning rate = 1.0, ρ = 0.95, $\varepsilon = 1 \times 10^{-6}$), mini-batch size 64, and up to 80 epochs with early stopping on validation loss (patience = 10). The CNN stream contains approximately 16.94 M trainable parameters in the 7-class setting.

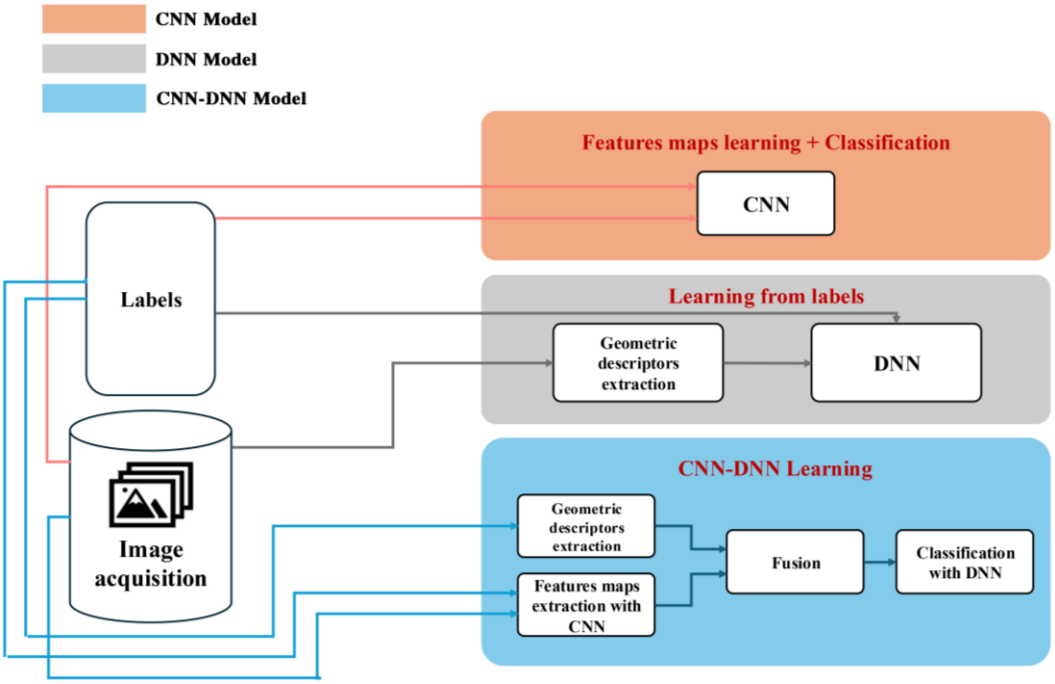

Fig. 2. General operation of CNN-DNN approach.

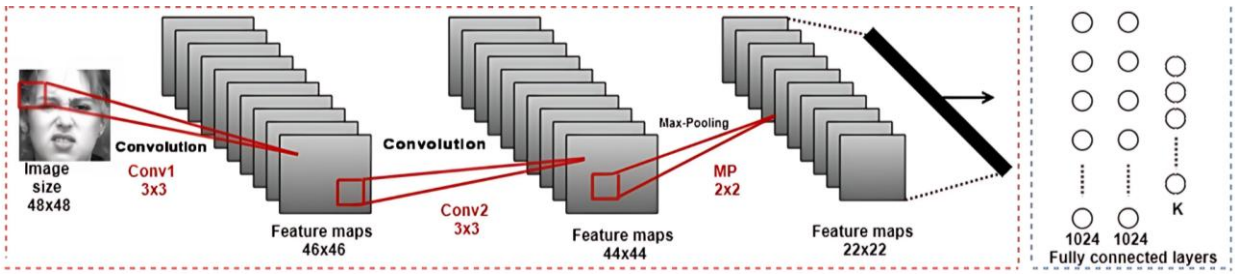

Fig. 3. Architecture of our CNN model.

3) Fully connected Deep Neural Network (DNN)

The DNN branch deals with geometry-based features that are computed from facial landmarks (Fig. 4). First, faces are detected via the Viola-Jones algorithm [36], and then they are resized to a fixed size of 256 × 256. The facial landmarks are then localized via the supervised descent method as proposed by Xiong and Torre [37]. The detected points are then paired to compute the Manhattan distance to obtain a distance descriptor of dimensionality 1176. To avoid redundancy and enhance efficiency, a Correlation-based Feature Selection (CFS) technique is employed to retain the most discriminative distances (71 features), which results in better accuracy and minimizes the risk of overfitting [38].

Finally, the selected features (71 distances) are fed to our DNN classifier for learning and classification. The DNN classifier, shown as the final block in Fig. 4, contains three hidden layers totaling 1024 neurons and uses ReLU activations. The output layer has K units (K = 6 or 7 depending on the dataset) and applies a SoftMax function to assign the input distance vector to one of the K emotion categories. For clarity, the three hidden layers are configured as 512, 256, and 256 units with ReLU activations; the resulting DNN contains approximately 0.236 M trainable parameters in the 7-class setting. Training follows the same optimization schedule as the CNN stream.

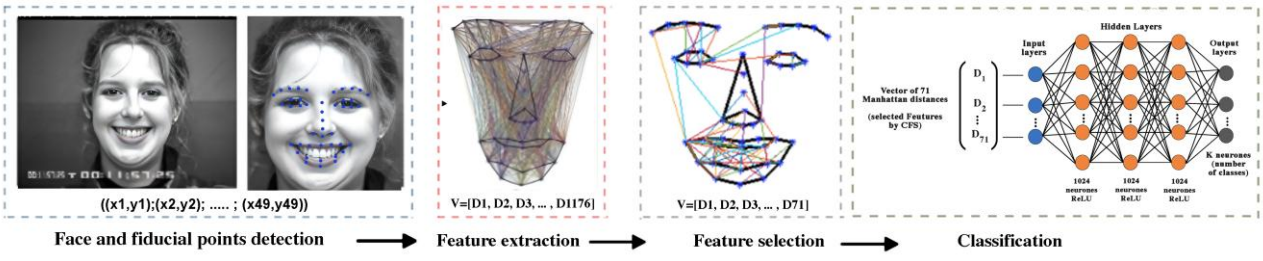

Fig. 4. An overview of the process followed for the DNN model.

4) Hybrid deep neural network (CNN-DNN)

The hybrid CNN-DNN model (Fig. 5) combines the previously described appearance and geometry branches into a single trainable network. By learning from pixel-level texture cues and landmark-based shape cues simultaneously, the fusion model is intended to improve recognition accuracy relative to either branch alone, while remaining lightweight enough for real-time use.

Concretely, the network ingests a 48 × 48 grayscale face crop for the CNN stream and a 71-dimensional vector of Manhattan distances for the DNN stream. Each stream produces an embedding through its respective layers (Figs. 3 and 4). The embeddings are then merged in the final classification stage, where a SoftMax layer outputs the predicted emotion label (Fig. 6). In our implementation, the CNN embedding (1024-D) and the DNN embedding (256-D) are concatenated and fed to a final SoftMax classifier (K = 6 or 7 depending on the dataset). The fused CNN-DNN model has approximately 17.18 M trainable parameters in the 7-class setting.

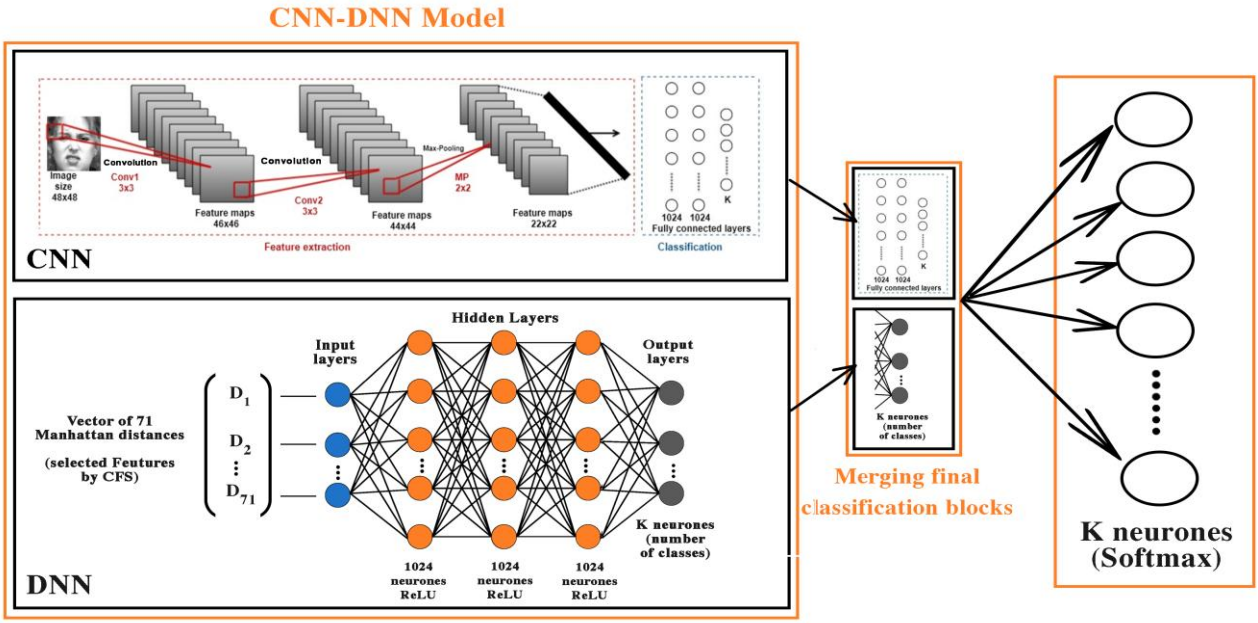

Fig. 5. An overview of the CNN-DNN model.

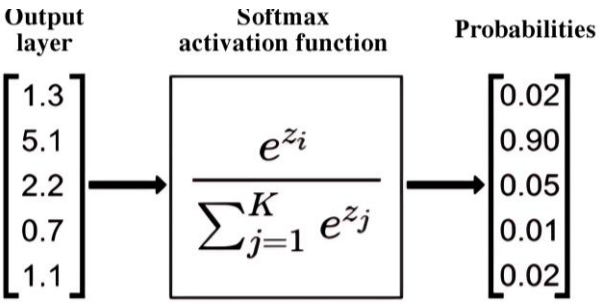

Fig. 6. An overview of the SoftMax activation function.

To instantiate the proposed architecture in a real institutional context, the platform was deployed as a FER-enabled VR learning environment for National School of Applied sciences (ENSA) of Khouribga, combining a modeled virtual campus, an interactive VR classroom and a web-based 360° video module.

The FER module outputs a probability distribution over emotion classes at each time step. In the next section, we describe how these outputs are exposed through a lightweight service interface and consumed by the VR and 360° clients to drive avatar facial animation and populate time-stamped affective analytics.

D. Integration into the VR Classroom

1) Virtual campus modeling

The virtual campus was modeled as a faithful replica of ENSA Khouribga using SketchUp, following the physical layout of buildings, corridors and outdoor spaces in order to support spatial orientation and a sense of place for students. The resulting 3D assets were exported in standard formats and organized into modular scenes (campus exterior, main hall, classrooms, laboratory areas) to facilitate integration into the Unity3D development environment and subsequent optimization for real-time rendering (Fig. 7).

Fig. 7. 3D design of the ENSA Khouribga using SketchUp.

The speed of development of digital technologies and the increasing need for adaptive teaching practices have driven the interest in the use of distance learning solutions. In this context, the challenge of mirroring the quality of interaction and engagement of the cognitive and emotional processes involved in face-to-face teaching has constituted a significant challenge. This has been especially the case concerning the ability of the traditional learning environment to create the feeling of presence and provide the non-verbal components of communication essential in teaching. In this regard, the development of our innovative immersive learning platform in the field of virtual reality and AI-based facial expressions presents itself as a remedy to the above challenges. The learning platform has the ability to break the boundaries of teaching through the provision of a rich educational experience rendered in a realistic representation of the National School of Applied Sciences of Khouribga through its AI-driven emotion analysis system based on deep learning. This allows the possibilities of immersion and interaction not only to be widened but also allows the teaching practices to be adapted according to the emotions of the learner.

The creation of a believable and recognizable virtual world becomes the cornerstone of our platform. The level of spatial rendering's fidelity becomes a crucial factor in its ability to induce the feeling of presence, along with the ability to correctly orient the user. In this regard, the project's implementation has been initiated through the completely contextual three-Dimensional (3D) modelling of the National School of Applied Sciences (ENSA) of Khouribga's physical environment. The software chosen has been SketchUp, which is also a Computer-Aided Design (CAD). The software gained preference due to its ease of learning, along with an intuitive interface and ability to work efficiently even when the geometries involved become complex. The software provides immense benefits due to its large libraries of models of pre-existing 3D items, along with the ability to create realistic photo-texture mapping, apart from being exportable in industry-standard formats such as Autodesk Filmbox (FBX) and Wavefront Object (OBJ). The software's ability to be exported in industry-standard formats such as FBX or OBJ helped immensely in the easy integration of the Unity 3D development platform used later.

2) VR classroom implementation

Based on these assets, an interactive multi-user VR classroom was implemented in Unity as the main pedagogical space. The scene contains a teacher avatar positioned at the front of the room, desks for students, presentation surfaces for slides and 3D objects, and interaction elements such as menus and teleportation anchors. The application supports navigation using VR controllers or gamepads, spatialized voice communication, and access to learning resources via in-world panels and menus (Figs. 8 and 9). A dedicated user interface layer exposes session control functions such as joining or leaving a session, enabling FER, and opening 360° content, which can be operated by instructors using standard VR headsets connected to a mid-range PC.

The move from static modelling to an immersive experience has been possible through the Unity 3D game development engine (Fig. 8). The reason why Unity has been chosen can be attributed to its robust nature and support from the Asset Store community for various development platforms across the globe. In addition to this, Unity has a lot of support from various VR headsets available in the market today, such as the Meta Oculus Quest. The support from the Asset Store community also allows us to integrate different software modules together effortlessly. This becomes an important factor in the context of incorporating our AI software modules.

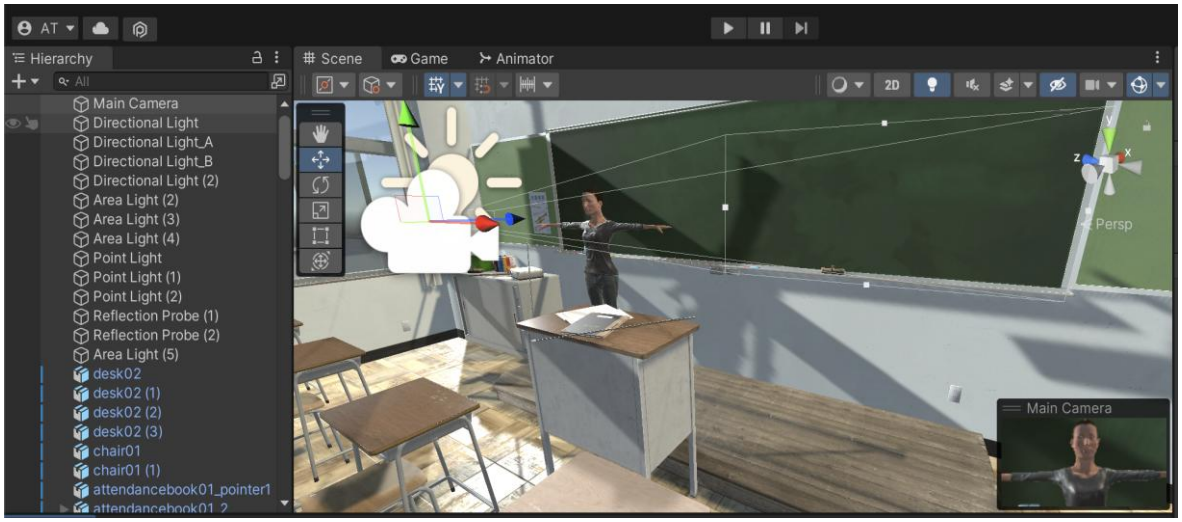

Fig. 8. Scene of a virtual classroom with avatar developed by unity engine.

The development work of the Unity platform involved the importation and setup of the 3D model of ENSA Khouribga. Certain enhancements were implemented to optimize the model's processing when rendered in the VR environment. The addition of the VR navigation system marked another crucial development phase of the

Unity platform. This provides users of the platform with the ability to move from location to location in the platform without developing motion sickness. The development of interaction systems marked another phase of this work. Users can interact with the different models used in the platform. Additionally, users can access the UI components of the learning platform.

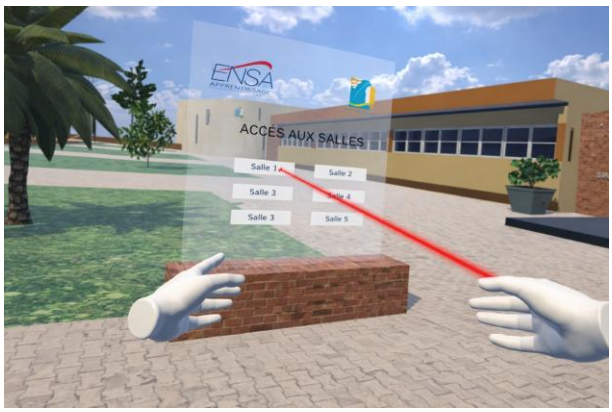

Fig. 9. Typical user interface in the VR environment, showing navigation and interaction options.

3) FER service integration and Application Programming Interface (API)

The FER module was integrated as an external service that receives video streams from user webcams, applies the CNN-DNN pipeline and returns high-level affective labels to the VR application. Video frames are first pre-processed using OpenCV-based face detection and landmark localization; grayscale face crops and normalized geometric descriptors are then fed to the trained CNN-DNN model, which outputs probability scores over the emotion classes considered in this study (e.g., joy, surprise, confusion, frustration, boredom, engagement), while relying on OpenCV for face detection and geometric pre-processing. The CNN branch was initialized by pre-training on large-scale FER datasets such as FER2013 and subsequently fine-tuned together with the geometric branch on the target databases used in this work, allowing the model to learn discriminative facial features for emotion classification.

In this paper, we use the term FER for the computer-vision task of predicting discrete expression categories from facial imagery. Public benchmark datasets used for training/evaluation largely follow basic-emotion taxonomies (e.g., happiness, anger, sadness, surprise, fear, disgust, neutral, and sometimes contempt). In the platform, we report expression labels to support awareness and reflection; when higher-level learning-relevant states (e.g., engagement, confusion, boredom) are discussed, they should be interpreted as coarse, application-level proxies rather than clinically valid affect labels. A careful mapping between dataset labels and learning-centered constructs requires additional validation and, ideally, multimodal evidence (e.g., voice and interaction signals).

Within Unity, a lightweight client library periodically queries the FER service through a RESTful API and maps the predicted emotions to avatar facial blend shapes and simple affective indicators displayed to the instructor. The Unity application captures facial images from the user's webcam (or VR headset camera, when available and permitted), sends them over a secure (Hypertext Transfer Protocol) (HTTP) request to the Python-based FER service, and receives the corresponding probability distribution over basic emotions. FER events are also logged in the affective database described in Section III-A, enabling the computation of temporal indicators such as the proportion of time spent in confusion during a given activity. At this stage, FER information was used primarily for awareness and reflection—for example, discreetly indicating dominant emotions on avatars, providing anonymous summaries of class-level engagement, or identifying segments with high confusion and prompting additional explanations—rather than for fully automatic adaptation of the learning scenario. The overall interaction between the Unity application and the FER service is illustrated in Fig. 10, and an example of the FER interface in Unity is shown in Fig. 11.

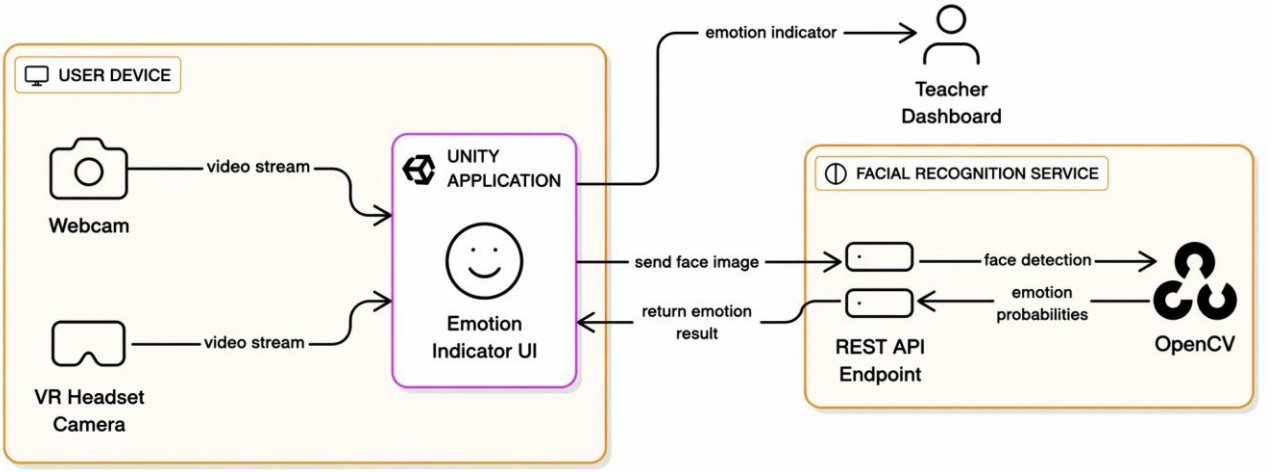

Fig. 10. Architectural diagram illustrates the interaction between the Unity application and the facial recognition service.

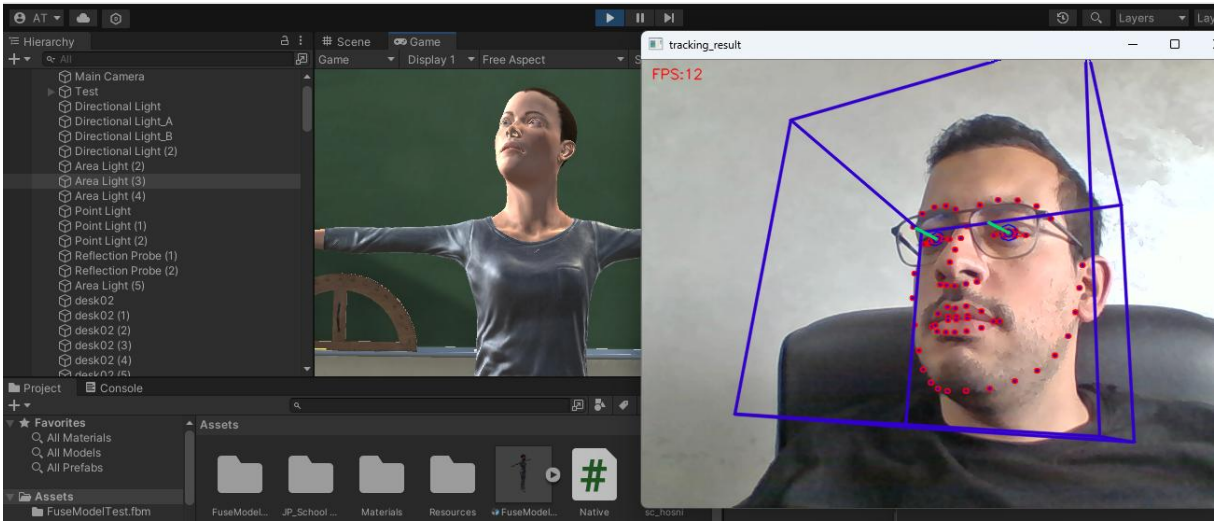

Fig. 11. Example interface showing facial expressions recognition using python library in Unity.

This rollout of features has significant ethical and privacy concerns. By further clarification, it is stated that the collection of facial data is done after obtaining consent from the user. The data is processed in a secure environment, with the option of anonymizing it for analysis. The underlying factor in this process is to enhance the learning experience.

In conclusion, this proposed solution allows for the creation of a VR platform that encompasses many elements that can create an engaging distance learning experience for the end-user.

The prototype combines spatial immersion through exploration of a faithful 3D replica of ENSA Khouribga with synchronous social interaction via avatars, chat, and spatialized voice. It also provides pedagogical access to virtual classrooms, resource consultation, and collaborative tools, while integrating real-time facial-expression analysis to support feedback and potential adaptation. Finally, the system is designed for hardware compatibility, supporting standard VR headsets (e.g., Meta Oculus Quest) and a potential desktop mode to increase accessibility.

While the VR classroom uses FER predictions primarily for real-time social signaling and instructor awareness, the same FER service also supports a lower-cost immersion pathway through 360° video. We therefore implement a parallel 360° learning mode in which learners access immersive scenes through a web player, while FER events are logged and summarized at the session level for later review.

4) Immersive alternative via 360° video

Being aware of the technological and economic constraints that VR might imply for users (need for a VR headset and a PC to support it, etc.), an alternative solution involving immersive video in 360° has also been investigated and has been developed. This allows a good compromise to be achieved between immersion and accessibility.

This involves filming real-school settings and class sessions using 360° cameras. The films are thereafter rendered (stitching and colour enhancement) through the aid of software applications (such as Premiere Pro from Adobe and Insta360 Studio). The films can also be equipped with interaction points in the form of hotspots, enabling users to transition from one scene to another (such as from the corridor to the classroom) or provide additional educational support materials (text information and images).

These are then used in the delivery of the immersive learning experience through the provision of videos that can be interacted with using a 360° effect (Fig. 12), available through an online platform using a browser, either through a computer or a mobile device. Though less advanced than the real-time version when it comes to interactivity support, the option presents a viable method of improving access to immersive learning through less expensive requirements from the learner's side when it comes to the support of the learning platform.

Fig. 12. Interface of the 360° video solution.

E. Experimental Settings and Implementation Details

With the platform components defined, we now describe how the system is evaluated. We separate the evaluation into two parts: first, an offline benchmark of the FER module on standard datasets to quantify recognition performance under controlled conditions; second, a pilot study that examines the feasibility and perceived pedagogical value of FER-enhanced immersive teaching in a realistic VR scenario.

1) Datasets and Training Protocol

a) Overview of evaluated FER approaches

We benchmarked three families of FER methods on CK+ [39], JAFFE [40], and OULU CASIA-VIS [41] to quantify the contribution of geometry, feature selection,

and deep fusion.

We benchmark three families of FER methods: geometry-based approaches, where facial landmarks are converted into distance-based descriptors (Euclidean, Minkowski, and Manhattan) to capture expression-related deformations; feature-selection variants that apply Correlation-based Feature Selection (CFS) to retain only the most informative distances and improve generalization; and deep learning approaches based on a hybrid CNN–DNN, where the CNN learns appearance features from face images, and the DNN learns complementary geometric patterns from landmark descriptors.

For the purely geometric baseline, we compute 1176 inter-point distances from facial landmarks and use Euclidean, Minkowski, and Manhattan variants to characterize expression-related deformation. These features are classified with conventional learners (Classification and Regression Tree (CART), Multilayer Perceptron (MLP), and Support Vector Machine (SVM) [42]) on both static images and dynamic sequences. Next, CFS reduces the 1176 distances to a compact set of 71 informative features to improve efficiency. For the deep model, the CNN-DNN fusion combines grayscale appearance information with the selected geometric vector. To increase robustness, we apply data augmentation to expand the training set by a factor of 16 using rotations, zoom, shifts, and horizontal flips. Finally, all datasets are split in a subject-independent manner (e.g., CK+: 92 subjects for training and 26 for testing) and we also report cross-database evaluation across different class settings (6 emotions, and 7-class variants including contempt or neutrality as applicable).

Limitations of benchmark datasets and domain shift: CK+, JAFFE, and OULU-CASIA are controlled datasets with predominantly posed expressions and relatively constrained capture conditions. Consequently, the reported benchmark accuracies should be interpreted as a controlled diagnostic of the model's capacity rather than as direct evidence of robustness in authentic educational settings. The cross-database drops observed in the cross-dataset evaluation illustrate domain shift; therefore, in the current platform we use FER outputs conservatively (awareness and aggregated analytics) and do not use them for high-stakes decisions. Future work will validate the model on in-the-wild FER datasets and on anonymized, consented in-platform, recordings with human annotation

b) Evaluation metrics

To compare methods rigorously, we report both predictive performance and resource requirements. Following common practice in FER, the selected metrics summarize classification behavior across classes and, for deep models, provide indicators of training convergence and deployment cost.

- Classification performance metrics

Performance metrics are computed from the confusion matrix on the validation/test sets. For each class (treated as the positive class in turn), we denote True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) to derive the following measures:

Accuracy: the overall fraction of correctly classified samples across all classes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision: for a given emotion class, the fraction of predicted positives that are correct (TP/(TP+FP)).

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall (Sensitivity): for a given emotion class, the fraction of true positives that are recovered (TP/(TP+FN)).

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-Score: the harmonic mean of precision and recall, providing a balanced indicator when class distributions are uneven.

$$F1 - Score = \frac{2 \times \left(Precision \times Recall \right)}{Precision + Recall} \tag{4}$$

Loss value: for deep models, we also report the validation loss (e.g., categorical cross-entropy) as an indicator of optimization progress and generalization.

- Model complexity metrics

To characterize computational cost for CNN, DNN, and CNN-DNN, we additionally report:

Number of parameters: the total count of trainable weights and biases, reflecting model capacity and complexity (CNN ≈ 16.94 M; DNN ≈ 0.236 M; fused CNN–DNN ≈ 17.18 M trainable parameters for the 7-class setting).

Model size: the storage footprint (MB) of the trained weights, which impacts deployment and loading time.

Together, these measures provide a concise but comprehensive basis for comparing accuracy, robustness, and practical deployability of the evaluated approaches.

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

2) Implementation details of the proposed VR application

To examine the pedagogical viability of the platform, we ran a user study centered on the proposed VR classroom with integrated facial-expression feedback. The experiment involved 65 university instructors (28 female, 37 male) aged 30–55 years (mean = 42.3) from the Higher Institute of Information and Communication (ISIC), Morocco. Participants came from varied teaching areas, including communication, humanities, politics, journalism, and related disciplines.

a) Proposed VR solution

The study used a Meta (Oculus) Quest 2 headset to deliver the immersive classroom experience, while the instructor's face was captured with a Sony Z90 camcorder and streamed via a Blackmagic Web Presenter to enable low-latency expression analysis. The VR classroom prototype was developed in Unity and configured to support the study workflow (Fig. 13).

Fig. 13. Demo of VR-classroom in experimental study.

b) Experimental setup

Each instructor delivered a short VR mini-lecture (5 min) to a simulated class of 10 AI-driven avatars. To reduce content-related confounds, all participants followed the same standardized prompt and timing, and the VR scene, avatar configuration, and feedback display were identical across sessions. During delivery, the instructor's facial video was captured live and processed by the FER module; the predicted affective state was then mapped onto the tutor's avatar to animate facial expressions. FER outputs were additionally logged for technical analysis (e.g., latency checks and qualitative inspection of obvious failure cases). We do not claim a formal, quantitative measure of emotional congruence because the avatar animation directly reflects the model prediction; instead, we report instructors' perceptions of expressiveness and note this as a limitation requiring a dedicated annotation-based validation study in future work.

Immediately after the session, participants completed a mixed-methods questionnaire to evaluate perceived usability, social/emotional presence, and technical quality (15 Likert items, 1 = strongly disagree to 5 = strongly agree), followed by five open-ended questions on barriers and improvement opportunities. The Likert items were adapted from commonly used usability and social-presence constructs in immersive learning research; the full item list is provided in Appendix A for transparency and reuse. Quantitative responses were analyzed using descriptive statistics and Pearson correlations (n = 65). Open-ended responses were analyzed using thematic coding by two researchers; disagreements were resolved through discussion to reach a consensus codebook. Given the exploratory nature of the pilot, qualitative findings are reported as indicative themes rather than as saturation claims.

IV. Results

A. FER Benchmark Results

On all geometric baselines, Manhattan distances produced the best results: 93.26% accuracy with SVM on CK+ sequences, and 78.57% accuracy with SVM on JAFFE static images. CFS further improved the accuracy: 100% accuracy on CK+ (6 classes) and 95.5% accuracy in a 7* setting with neutrality. The proposed CNN-DNN fusion technique produced the highest accuracy: 100% accuracy on CK+ and 81.25% accuracy on OULU, whereas cross-dataset testing remains challenging on JAFFE (47.89%) as reported in Table I. Figs. 14–16 present classification rates of CNN, DNN, and CNN-DNN techniques on all datasets.

TABLE I. 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%

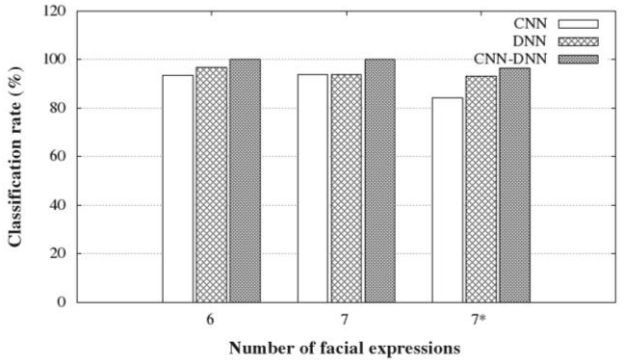

Fig. 14. Classification accuracy (%) of CNN, DNN, and CNN-DNN models on the CK+ database.

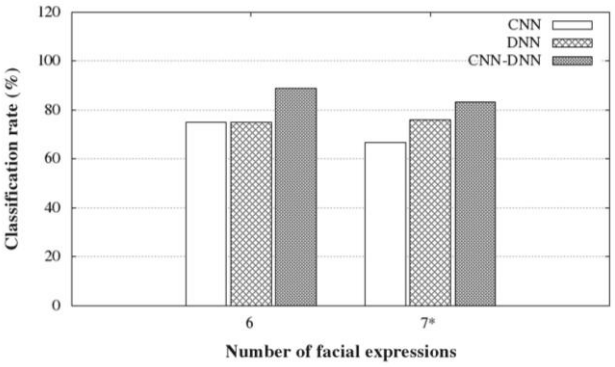

Fig. 15. Classification accuracy (%) of CNN, DNN, and CNN-DNN models on the JAFFE database.

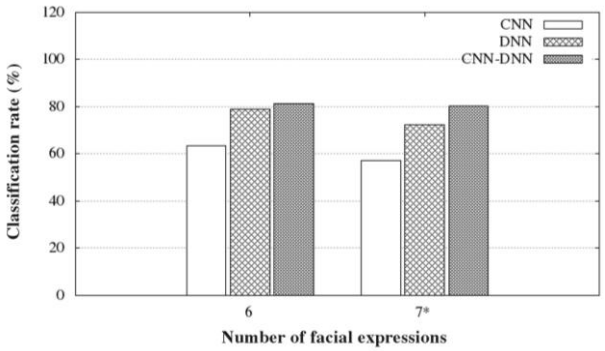

Fig. 16. Classification accuracy (%) of CNN, DNN, and CNN-DNN models on the OULU-CASIA database.

B. Cross-Dataset Results

In cross-database experiments, the domain shift problem of FER is emphasized. Good generalization performance was obtained from OULU to CK+ (up to 90.89%), which might be due to some similarity in capture conditions and expression variation. However, the performance declined significantly when testing on JAFFE (e.g., 53.55%) as presented in Table II, which may result from differences in demographics, illumination, and expression types of posed/spontaneous.

We also observed that adding the extended class configurations (7* or 7 classes) tends to reduce accuracy compared with the 6-class setting. This degradation can be plausibly related to differences in finer-grained categories (neutrality vs. subtle micro-expressions) as well as class imbalance and visual similarities among some emotions.

TABLE II. CROSS-DATASET EVALUATION

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

In conclusion, these baselines reinforce the notion that temporal information can be useful, with expression evaluation using sequences making the geometric approach more informative, particularly with Manhattan Distance, which is robust to inter-subject variation. CFS has an important double effect: it improves accuracy up to 100% on CK+ while reducing the dimensionality of the features from 1176 to 71, thus reducing the computational cost and mitigating overfitting.

Lastly, the best accuracy is achieved by the deep fusion model in most cases because it combines texture/appearance features from the CNN with other shape features from the DNN, which are complementary to texture features. The weak spots, particularly with JAFFE, likely reflect the importance of dataset size, diversity, and collection conditions for FER.

C. VR/360 Pilot Study Results

Across the VR teaching trials, the combined quantitative scores and open-ended feedback indicated distinct patterns of adoption and integration readiness. We summarize the main observations in the following themes (acceptance, barriers, technical performance, and correlation analysis).

1) Acceptance of VR

A comparative bar-chart summary of these presence indicators is provided in Fig. 17 to make the improvement more explicit.

Table III summarizes instructors' acceptance of the VR environment. It can be seen that there were more positive views on the environment than negative, with notable focus on the ability to effectively communicate emotion using the environment, as well as recognizing learners' confusion and involvement through avatars.

2) Adoption barriers

According to Table IV, the primary challenges facing the implementation of virtual reality technology within the educational sector include the challenge of implementation, high costs of hardware, infrastructure problems, and a lack of user preparation. Most of the teachers interviewed noted that implementing virtual reality within their curriculum is difficult, with the major problem being the cost factor and infrastructure, while a few had adequate skills to use the technology.

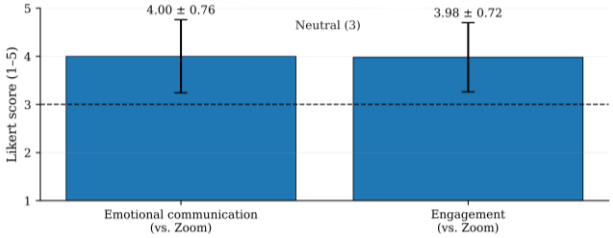

Fig. 17. Comparative visualization of survey-reported social/emotional presence improvements (VR platform vs videoconferencing baseline) using agreement rates and mean Likert scores ($N = 65$).

TABLE III. INSTRUCTORS' ACCEPTANCE AND PERCEPTION OF VR TECHNOLOGY

Aspect	Agreement (%)	Mean Likert Score	Key insights
Perceived support for emotional communication in VR	74	4.2	Reported as subjective perception (single-condition pilot; no controlled videoconferencing baseline).
Perceived support for engagement via avatar facial cues	72	4.1	Participants reported that avatar expressions helped notice confusion; qualitative themes highlighted confusion cues.

Note: "Agreement" denotes the percentage of respondents selecting Agree or Strongly Agree; "Mean Likert Score" is the average on a 1–5 scale (1 = strongly disagree, 5 = strongly agree).

TABLE IV. 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.

3) *Technical performance*

Regarding technical performance, 63% of respondents (mean = 3.1) characterized facial-expression detection as moderate. The most frequently reported failure mode was confusion between sadness and neutrality, cited by 42% of participants. A summary of these real-time indicators is provided in Fig. 18.

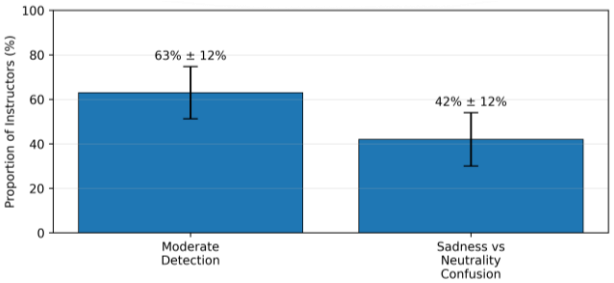

Fig. 18. Perceived real-time FER performance in the integrated platform: proportion of respondents rating detection as moderate and the most frequently reported confusion mode (sadness vs neutrality).

Despite these limitations, 81% of instructors highlighted the low-cost nature of the prototype (approximately a few hundred USD in commodity components) compared with commercial affect-sensing solutions (e.g., Affectiva), suggesting that the approach is suitable for pilot deployments in resource-constrained contexts.

4) *Key correlations*

Correlation analysis (n = 65) showed a positive association between age and resistance to adopting VR (r = 0.62, 95% CI [0.44, 0.75], $p < 0.01$), consistent with prior observations that older instructors may adopt immersive tools more cautiously.

We also found a negative correlation between prior technology experience and criticism of system performance (r = −0.45, 95% CI [−0.63, −0.23], $p < 0.05$), suggesting that less experienced users were generally more tolerant of early-stage imperfections. These correlations are exploratory and should be interpreted cautiously.

V. DISCUSSION

The results reported above can be interpreted at two complementary levels. First, the benchmark experiments clarify the technical strengths and current limits of the proposed FER pipeline under controlled evaluation conditions. Second, the pilot deployment helps explain how instructors perceive the pedagogical value, practical constraints, and governance conditions associated with FER-enabled immersive teaching. This section therefore, discusses the findings in terms of model performance and generalization, educational usefulness, deployment feasibility, and responsible institutional adoption.

A. *Interpretation of FER Results*

From the quantitative assessment, the geometric methods solely based on Manhattan distances between facial points can produce competitive results on this task and are further improved when dynamic information is fed into the analysis. Even when operating in an optimized 71-dimensional space of distances obtained via the application of Correlation Feature Selection (CFS), which reduces the original 1176 descriptors, it is possible to achieve 100% accuracy on CK+, six emotion classes, in certain scenarios.

Further improvement in performance for many settings is achieved due to the joint use of appearance and geometric cues in the hybrid CNN-DNN approach. In CK+ and University of Oulu-Institute of Automation, Chinese Academy of Sciences (OULU-CASIA), the proposed network reaches state-of-the-art accuracy in comparison to or better than many recent hybrid approaches in this area [37]. However, cross-setting evaluations state the problem of bias in training on only one collection or corpus: frameworks are acceptable for CK+, yet drastically degrade on validation sets for JAFFE, which may be depicted in changes in demographics and imaging conditions. Consistent with past findings concerning bias and domain gap in FER tasks [36], this makes it evident that, despite the valid potential of CNN-DNN networks, in practical applications in education, for instance, further approaches such as multi-corpus training and domain adaptation, and perhaps local fine-tuning of the network on self-acquired data may be mandatory.

Another observation is made concerning the accuracy-degrading effect of adding “neutral” or “contempt” to the emotion set (7 and 7* emotion sets). It is well known that such emotion labels are ambiguous and less represented in many sets, thereby leading to increased confusion among nearby labels. In the light of distance learning frameworks, this means that FER answers should rather be regarded as probabilistic signals of engagement or affective tendencies rather than definite answers for individual learners.

Bridging benchmark evaluation to platform deployment: the FER component is validated here on

public benchmarks to enable transparent comparison under controlled conditions; however, we explicitly acknowledge that this does not fully represent real classroom webcam footage, where lighting, head pose, occlusions, and spontaneous expressions are more variable. In the deployed prototype, FER predictions are used primarily to animate the instructor's avatar and to produce coarse, time-aggregated indicators (e.g., peaks of confusion) for reflection. A dedicated in-platform validation study—collecting consented recordings, obtaining human labels on educational segments, and reporting agreement between model predictions and annotations—is planned as the next step before any claims of real-world FER performance.

B. Implications for VR and 360° E-Learning

Results obtained from the pilot experiment conducted on 65 instructors indicate that the integration of Facial Expression Recognition (FER) in Virtual Reality (VR)-aided learning was perceived by participants as supporting communication. In particular, many instructors reported that—relative to their usual videoconferencing experience—the VR scenario felt more supportive of emotional communication; however, this reflects subjective self-report in a single-condition pilot and does not constitute an empirical baseline comparison. Participants also indicated that avatars' facial expressions helped them notice confusion and interest cues in the virtual class. This aligns with prior work suggesting that well-structured immersion in virtual reality can enhance perceived social presence and learning.

Simultaneously, the role of the 360° video component becomes equally supportive and valuable. It represents an easier on-ramp for immersive learning, in that it only requires access to a web browser and potentially an affordable headset. For those in lower-resource environments, what may be the only feasible approach is to supplement immersion in specific contexts with immersion that comes from a 360° video platform. In this paper, we clarify that the current study does not provide a dedicated user evaluation of the 360° module; it is included to demonstrate architectural integration, and a full learning-outcome and usability assessment is part of future work.

In terms of implications for design, what this means is that it is better for FER data to be included in analytics and feedback systems rather than in the form of labels of raw emotions. For instance, analytics will be of great use in representing the spread of expected emotional states and redesigning learning activities or planning breaks in line with this in VR lectures.

C. Adoption Barriers and Practical Constraints

Despite the aspirational mindset, some possible barriers for adoption were also identified. Instructors indicated that the cost of virtual reality hardware and the absence of sufficient network capacity were core barriers, especially in the case of public universities. Very few instructors have reported experience in the use of virtual reality software, establishing the necessity for special training and assistance in this area. Results are in line with existing literature on making education mainstream in virtual reality, especially concerning the cost of hardware and network maintenance [16].

"Acceptable" levels of technical performance were reported for the majority, though misclassifications, like the pairing of sadness and neutral expressions, were observed. In this case, the impact of such misclassifications may be of specific pedagogical importance in the classroom, given that a misinterpretation of affect can affect teaching outcomes. Significantly, too, is the finding showing that a strong link between the age of the instructors and resistance to the use of VR" exists.

Finally, the implications of affective computing and FER in relation to ethics and the right to privacy must be carefully considered. Even if the process involves local and anonymous data processing, the aspect of analyzing faces might be seen as invading the privacy of some students or lecturers.

D. Toward Affect-Aware Immersive Learning

Overall, the findings support the feasibility of integrating deep learning-based FER as a core service within low-cost VR and 360° e-learning platforms. The hybrid CNN-DNN approach offers a promising trade-off between accuracy and computational cost on a mid-range personal computer, and the proposed platform demonstrates the application of FER in a way that can enhance social presence and deliver lightweight affective analytics.

Nevertheless, the study highlights the need for a cautious approach in interpretation and usage. In practice, FER should be used as one of the contributing factors, along with other sources, and should not be used in isolation as a so-called "emotion detector". Thus, in line with the findings, the study of affect-aware immersive learning environments should progress in the directions of multi-modal sensing, fair and robust modeling in diverse cultures, and teacher-in-the-loop adaptation, where the teacher remains in control of the timing and application of affective analytics in teaching.

E. Ethical and Privacy Considerations

Continuous or regular facial monitoring within an educational context gives rise to several legitimate concerns about privacy, informed consent, power relationships, and emotional surveillance. Accordingly, the proposed platform should be viewed as a research tool to be tested only with informed consent. Several practical measures can be implemented, such as opt-in participation with an opt-out option that does not restrict access to learning materials, data minimization (e.g., processing frames on a server or institutionally, rather than storing raw video), purpose restriction, disclosure of what is inferred, confidence, and use, retention restriction, safe handling of logs if collected, and testing for demographic bias and domain shift before use. Based on the misclassifications noted by participants, the results

of facial emotion recognition should not be used for any consequential purposes, such as grading or disciplinary action.

F. Limitations of the Pilot Study

Participants were drawn from a single institution and comprised instructors only; therefore, the findings reflect an exploratory, context-specific perspective and should not be generalized to other universities, disciplines, or student populations. Future work will include multi-institution samples, student participants, and controlled comparisons in authentic online courses.

VI. Conclusion

The research aims to examine the viability of integrating immersive learning delivery with affect-aware analytics in one system. Two key findings are noted from this research. From a technical standpoint, the hybrid CNN-DNN model shows strong performance in benchmark testing, with 100% accuracy in CK+ and 81.25% accuracy in OULU-CASIA in 6-class classification, affirming the benefits of integrating both modalities in a subject-independent framework. From a cross-dataset standpoint, while there is evidence of constrained generalization in varying acquisition and user populations, a cautious perspective is advised in interpreting FER outcomes, which should be viewed as advisory signals rather than determinative emotional states.

From an application standpoint, while a pilot study of 65 instructors found promise in enhancing social and emotional communication in immersive learning delivery, there was a clear identification of considerable barriers to implementation, such as hardware costs, infrastructure, and training, alongside concerns over misclassifications, which must be overcome prior to wider-scale implementation in an institutional context.

From a contribution standpoint, this research contributes to existing literature in three ways: theoretically, by integrating virtual reality, web-based 360 learning, and FER-enabled affective feedback in one system, rather than addressing them in siloed ways, and practically, by providing a relatively low-cost model for institutions unable to afford expensive commercial affect-sensing solutions, while still allowing for future evolution and expansion, and ethically, by emphasizing the importance of informed consent, purpose, and control over participation in FER-enabled scenarios.

Overall, the contribution of the paper lies less in claiming a fully validated emotion-aware classroom than in demonstrating a feasible and carefully delimited foundation for responsible affect-aware immersive learning. The results show that technical performance, teacher acceptability, deployment feasibility, and governance safeguards must be considered together if such systems are to be educationally credible and socially acceptable.

Future work will therefore focus on improving robustness with more diverse and in-the-wild data, validating in-platform FER behavior through human annotation and multimodal evidence, extending evaluation to students and learning outcomes under controlled comparison conditions, and designing privacy-preserving, fairness-aware, teacher-in-the-loop dashboards that support pedagogical decision-making without automating high-stakes judgments.

Appendix A Pilot Study Questionnaire (Likert Items)

Participants rated each statement on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree).

Note: item wording is provided for transparency; future studies will further validate the instrument and report reliability.

(1) The VR classroom interface was easy to learn and operate.
(2) I could focus on teaching without being distracted by the controls.
(3) Audio/voice communication in the VR classroom was clear enough for teaching.
(4) The VR classroom gave me a sense of "being together" with learners compared to typical online teaching.
(5) I felt more socially connected to the audience in the VR classroom.
(6) Seeing the avatar facial expressions helped me interpret learners' reactions.
(7) FER-driven avatar expressions helped me notice confusion or lack of understanding.
(8) FER-driven feedback could help me adapt my teaching pace or explanations.
(9) The avatar's facial animation appeared sufficiently natural for instructional use.
(10) Misclassifications or unnatural expressions negatively affected my confidence in the system.
(11) I would be willing to use this type of VR classroom for certain course activities.
(12) The required equipment and setup would be feasible in my institution.
(13) I have privacy concerns about continuous facial expression monitoring in teaching/learning settings.
(14) The platform should allow users to participate without enabling FER if they prefer.
(15) Overall, this platform concept is valuable for improving engagement and communication in distance education.

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.

REFERENCES

[1] D. R. Garrison, T. Anderson, and W. Archer, "Critical inquiry in a text-based environment: Computer conferencing in higher education," *The Internet and Higher Education*, vol. 2, no. 2–3, pp. 87–105, 2000. doi: 10.1016/S1096-7516(00)00016-6

[2] J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt, "A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda," *Computers & Education*, vol. 147, 103778, 2020. doi: 10.1016/j.compedu.2019.103778

[3] L. Jensen and F. Konradsen, "A review of the use of virtual reality head-mounted displays in education and training," *Education and Information Technologies*, vol. 23, no. 4, pp. 1515–1529, 2018. doi: 10.1007/s10639-017-9676-0

[4] L. E. 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," *Journal of Xi'an Shiyou University, Natural Sciences Edition*, vol. 67, no. 6, 2024. doi: 10.5281/zenodo.11543356

[5] K. Kawashima, F. Nader, J. W. Collins, and A. Esmaeili, "Virtual reality simulations in robotic surgery training: a systematic review and meta-analysis," *Journal of Robotic Surgery*, vol. 19, no. 1, 29, 2024. doi: 10.1007/s11701-024-02187-z

[6] Y. Zhang, Y. Wang, G. Fan, Y. Song, and Y. Hu, "Multimodal teaching analytics: The application of SCORM courseware technology integrating 360-degree panoramic VR in historical courses," *Scientific Reports*, vol. 13, 18893, 2023. doi: 10.1038/s41598-023-46229-2

[7] A. Kopalidis, E. Mouchlianitis, and N. Nikolaidis, "Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets," *Information*, vol. 15, no. 3, 135, 2024. doi: 10.3390/info15030135

[8] N. Sun, Y. Song, J. Liu, L. Chai, and H. Sun, "Appearance and geometry transformer for facial expression recognition in the wild," *Computers and Electrical Engineering*, vol. 107, 108583, 2023. doi: 10.1016/j.compeleceng.2023.108583

[9] Z. Zhang, J. M. Fort, and L. G. Mateu, "Facial expression recognition in virtual reality environments: Challenges and opportunities," *Frontiers in Psychology*, vol. 14, 1280136, 2023. doi: 10.3389/fpsyg.2023.1280136

[10] J. Kim, H. Lee, and S. Park, "Facial expression recognition in the wild using face graph and attention," *IEEE Access*, vol. 11, pp. 59774–59787, 2023. doi: 10.1109/ACCESS.2023.3286547

[11] A. Touima and M. Moughit, "Facial expressions in virtual reality based-education: Understanding recognition approaches and their integration for immersive experiences," *Journal of Image and Graphics*, vol. 13, no. 6, pp. 604–620, 2025. doi: 10.18178/joig.13.6.604-620

[12] D. Burke, H. Crompton, and C. Nickel, "The use of Extended Reality (XR) in higher education: A systematic review," *TechTrends*, vol. 69, pp. 998–1011, 2025. doi: 10.1007/s11528-025-01092-y

[13] X. Han, H. Luo, Z. Wang, and D. Zhang, "Using virtual reality for teacher education: A systematic review and meta-analysis of literature from 2014 to 2024," *Frontiers in Virtual Reality*, vol. 6, 1620905, 2025. doi: 10.3389/frvir.2025.1620905

[14] N. L. Schroeder, R. F. Siegle, and S. D. Craig, "A meta-analysis on learning from 360° video," *Computers & Education*, vol. 206, 104901, 2023. doi: 10.1016/j.compedu.2023.104901

[15] Z. Jiang, Y. Zhang, and F. K. Chiang, "Meta-analysis of the effect of 360-degree videos on students' learning outcomes and non-cognitive outcomes," *British Journal of Educational Technology*, vol. 55, no. 6, pp. 2423–2456, 2024. doi: 10.1111/bjet.13464

[16] Mohd 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," *Sustainable Futures*, vol. 8, 100324, 2024. doi: 10.1016/j.sftr.2024.100324

[17] G. Makransky and R. E. Mayer, "Benefits of taking a virtual field trip in immersive virtual reality: Evidence for the immersion principle in multimedia learning," *Educational Psychology Review*, vol. 34, pp. 1771–1798, 2022. doi: 10.1007/s10648-022-09675-4

[18] G. Lampropoulos and Kinshuk, "Virtual reality and gamification in education: A systematic review," *Educational Technology Research and Development*, vol. 72, pp. 1691–1785, 2024. doi: 10.1007/s11423-024-10351-3

[19] S. Luo, D. Zou, and L. Kohnke, "A systematic review of research on xReality (XR) in the English classroom: Trends, research areas, benefits, and challenges," *Computers & Education: X Reality*, vol. 4, no. 100049, 2024. doi: 10.1016/j.cexr.2023.100049

[20] T. Monahan, G. McArdle, and M. Bertolotto, "Virtual reality for collaborative e-learning," *Computers & Education*, vol. 50, no. 4, pp. 1339–1353, 2008. doi: 10.1016/j.compedu.2006.12.008

[21] Y. Tian, J. Zhu, H. Yao, and D. Chen, "Facial expression recognition based on vision transformer with hybrid local attention," *Applied Sciences*, vol. 14, no. 15, 6471, 2024. doi: 10.3390/app14156471

[22] J. C. Kim, M. H. Kim, H. E. Suh, M. T. Naseem, and C. S. Lee, "Hybrid approach for facial expression recognition using convolutional neural networks and SVM," *Applied Sciences*, vol. 12, no. 11, 5493, 2022. doi: 10.3390/app12115493

[23] M. Aly, "Revolutionizing online education: Advanced facial expression recognition for real-time student progress tracking via deep learning model," *Multimedia Tools and Applications*, vol. 84, pp. 12575–12614, 2025. doi: 10.1007/s11042-024-19392-5

[24] K. Shingjergji, D. Iren, C. Urlings, and R. Klemke, "Affective computing in online higher education: A systematic literature review," *Computers and Education: Artificial Intelligence*, vol. 10, 100499, 2025. doi: 10.1016/j.caeai.2025.100499

[25] M. A. Vicente-Querol, A. Fernández-Caballero, J. P. Molina *et al*., "Facial affect recognition in immersive virtual reality: Where is the participant looking?" *International Journal of Neural Systems*, vol. 32, no. 7, 2250029, 2022. doi: 10.1142/S0129065722500290

[26] M. P. A. Ramaswamy and S. Palaniswamy, "Multimodal emotion recognition: A comprehensive review, trends, and challenges," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, no. 6, e1563, 2024. doi: 10.1002/widm.1563

[27] S. Yu, A. Androsov, H. Yan, and Y. Chen, "Bridging computer and education sciences: A systematic review of automated emotion recognition in online learning environments," *Computers & Education*, vol. 220, 105111, 2024. doi: 10.1016/j.compedu.2024.105111

[28] M. N. Hasnine, H. T. Nguyen, T. T. T. Tran, H. T. T. Bui, G. Akçapınar, and H. Ueda, "A real-time learning analytics dashboard for automatic detection of online learners' affective states," *Sensors*, vol. 23, no. 9, 4243, 2023. doi: 10.3390/s23094243

[29] S. Stöckli, M. Schulte-Mecklenbeck, S. Borer, and A. C. Samson, "Facial expression analysis with AFFDEX and FACET: A validation study," *Behavior Research Methods*, vol. 50, no. 4, pp. 1446–1460, 2018. doi: 10.3758/s13428-017-0996-1

[30] I. Domínguez-Catena, D. Paternain, and M. Galar, "Metrics for dataset demographic bias: A case study on facial expression recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5209–5226, 2024. doi: 10.1109/TPAMI.2024.3361979

[31] H. Li, Y. Luo, T. Gu, and L. Chang, "BFFN: A novel balanced feature fusion network for fair facial expression recognition," *Engineering Applications of Artificial Intelligence*, vol. 138, 109277, 2024. doi: 10.1016/j.engappai.2024.109277

[32] I. Dominguez-Catena, D. Paternain, A. Jurio, and M. Galar, "Less can be more: Representational vs. stereotypical gender bias in facial expression recognition," *Progress in Artificial Intelligence*, vol. 14, pp. 11–31, 2025. doi: 10.1007/s13748-024-00345-w

[33] Y. Zeng, Z. Zhang, J. Liu, J. Ma, and Z. Liu, "Pri-EMO: A universal perturbation method for privacy preserving facial emotion recognition," *Journal of Information and Intelligence*, vol. 1, no. 4, pp. 330–340, 2023. doi: 10.1016/j.jiixd.2023.08.001

[34] J. Zhang, X. Xie, G. Peng, L. Liu, H. Yang, R. Guo, J. Cao, and J. Yang, "A real-time and privacy-preserving facial expression recognition system using an AI-powered microcontroller," *Electronics*, vol. 13, no. 14, 2791, 2024. doi: 10.3390/electronics13142791

[35] A. Salman and C. Busso, "Privacy-preserving personalization for video-based facial expression recognition using federated learning," in *Proc. 24th ACM Int. Conf. on Multimodal Interaction (ICMI)*, 2022, pp. 548–556. doi: 10.1145/3536221.3556614

[36] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004. doi: 10.1023/B:VISI.0000013087.49260.fb

[37] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 532–539. doi: 10.1109/CVPR.2013.75

[38] M. A. Hall, "Correlation-based Feature Selection for Machine Learning," Ph.D. dissertation, Dept. Comput. Sci., Univ. of Waikato, Hamilton, New Zealand, 1999.

[39] 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. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 94–101, 2010, doi: 10.1109/CVPRW.2010.5543262.

[40] M. J. Lyons, M. Kamachi, and J. Gyoba. (2019). The Japanese Female Facial Expression (JAFFE) Dataset, *Zenodo*. [Online]. Available: https://zenodo.org/records/3451524.

[41] G. Zhao, X. Huang, T. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011. doi: 10.1016/j.imavis.2011.07.002

[42] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009. doi: 10.1007/978-0-387-84858-7