IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 3 February 2025, accepted 25 February 2025, date of publication 28 February 2025, date of current version 13 March 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3546772

RESEARCH ARTICLE

Multimodal Emotional Detection System for Virtual Educational Environments: Integration Into Microsoft Teams to Improve Student Engagement

WILLIAM VILLEGAS-CH[1], **(Member, IEEE), ROMMEL GUTIERREZ**[1], **AND ARACELY MERA-NAVARRETE**[2]

[1]Escuela de Ingeniería en Ciberseguridad, FICA, Universidad de Las Américas, Quito 170125, Ecuador
[2]Departamento de Sistemas, Universidad Internacional del Ecuador, Quito 170411, Ecuador

Corresponding author: William Villegas-Ch (william.villegas@udla.edu.ec)

ABSTRACT In modern educational settings, especially virtual environments, understanding students' emotional states has become crucial in improving participation, engagement, and academic performance. Current emotional detection systems, such as Affectiva and Emotient, offer facial and vocal emotional recognition but present limitations regarding integration, scalability, and adaptability to widely used platforms such as Microsoft Teams. These systems often require complex integrations and are not readily adaptable to the dynamic nature of the educational environment. This study proposes an innovative solution for real-time emotional detection, using a multimodal approach that combines speech and facial expression analysis within the Microsoft Teams platform. By integrating TensorFlow, PyTorch, OpenCV, and Dlib, we developed a system capable of accurately detecting emotions such as happiness, stress, and calm, with up to 95% detection precision for positive emotions. The system was tested in an educational environment, demonstrating its ability to process multiple interactions simultaneously without compromising performance. Results include high precision in essential emotion detection and significant improvements in student engagement. However, the system showed limitations in detecting more complex emotions such as stress and frustration, suggesting further refining the model.

INDEX TERMS Emotional detection, virtual education, Microsoft Teams, multimodal analysis.

I. INTRODUCTION

Emotional detection in the educational context has gained increasing relevance in recent years due to its ability to improve the interaction between students and teachers, thus optimizing the teaching-learning process [1]. Emotional recognition enables a deeper understanding of how students' emotions impact their participation, motivation, and performance in educational activities [2]. In online education, where face-to-face interactions are limited, understanding students' emotions becomes even more crucial. Emotions such as happiness, stress, or frustration significantly influence how students engage with content, participate in discussions, or manage their cognitive load during class [3]. However, accurately detecting these emotions in educational settings remains challenging due to the different forms of emotional expression and the necessity of integrating data from diverse sources, such as voice and facial expressions.

The proposed study introduces an innovative solution for emotional detection in virtual educational environments.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhe Jin.

The developed system integrates advanced acoustic signal processing techniques and facial image analysis through neural networks, leveraging Microsoft Teams as an academic platform for implementation [4]. This integration enables the detection of students' emotions during classes, providing teachers with valuable information to adjust pedagogical strategies and enhance student participation and motivation [5]. Multimodal detection, which fuses voice and facial expressions, has been demonstrated to offer superior accuracy over unimodal methods by addressing complex emotions that are not always evident in a single data source [6].

This work contributes a novel solution by directly integrating multimodal emotional detection into Microsoft Teams, addressing the critical limitations of existing tools such as Affectiva and Emotient [7]. While these tools achieve high precision in recognizing basic emotions, they require extensive technical adaptations and lack scalability for large virtual classrooms. In contrast, the proposed system achieves comparable precision while providing seamless integration, greater adaptability, and scalability. This design enables the application of emotional detection across diverse educational contexts without requiring specialized technical expertise or additional resources.

The justification for this work lies in the growing necessity for effective and accessible systems to monitor emotions in the classroom, particularly in virtual environments. While providing an indispensable response to global circumstances, these environments present unique challenges in fostering emotional connections between students and teachers [8], [9]. Platforms like Microsoft Teams, widely adopted in modern education, lack integrated tools for measuring and responding to participants' emotions. This limitation can lead to reduced interaction and diminished educational quality. Existing emotional detection systems, such as Affectiva and Emotient, often require custom integrations and lack the agility to adapt to widely used platforms like Teams, restricting their applicability in educational domains.

The methodology developed in this study combines Convolutional Neural Networks (CNN) for facial expression analysis and Recurrent Neural Networks (RNN), specifically Long Short-Term Memory (LSTM) models, for processing acoustic signals [12], [13]. These techniques were selected for their proven ability to capture temporal and spatial patterns in data, making them particularly suitable for emotion classification in voice and face sequences. TensorFlow and PyTorch were employed for voice data processing, while facial analysis utilized OpenCV and Dlib to precisely detect facial key points associated with emotions such as happiness, stress, and sadness. Datasets such as RAVDESS for vocal emotions and AffectNet for facial expressions were used to train these models, providing a robust foundation [14], [15].

The integration with Microsoft Teams was achieved using the Microsoft Graph API [16] and Azure Cognitive Services [17]. This integration facilitates real-time emotional detection during class sessions, presenting the results as actionable insights for teachers. The alerts generated by the system allow educators to adapt their teaching strategies in real-time, responding proactively to students' emotional needs. This aspect is particularly critical in fostering dynamic and inclusive classroom environments, where early intervention can mitigate disengagement and improve collaborative learning outcomes.

The main results of this study demonstrate that the system achieves high accuracy in detecting common emotions such as happiness and calm, with precision values reaching 0.95 for positive and neutral feelings. However, complex emotions such as stress and frustration presented challenges, with slightly lower detection scores, emphasizing the need for continued model refinement. The system also demonstrated scalability by processing multiple simultaneous interactions with low latency, making it suitable for large virtual classrooms. Its seamless integration with Microsoft Teams provided a simple and efficient implementation process without requiring specialized technical expertise, a significant advantage over other systems requiring more complex adjustments.

This study also highlights the unique contributions of this system in addressing the technical and pedagogical challenges of virtual education. The system ensures robust performance even in resource-constrained environments by integrating multimodal data fusion and cloud-edge processing. Additionally, the scalability and adaptability of this system enable its application across diverse educational contexts, from small workshops to large-scale institutional deployments.

This work innovates emotional education by delivering a scalable and practical solution for real-time emotional feedback. This enables teachers to enhance student engagement and motivation, key factors directly impacting academic performance. The multimodal approach captures a broader emotional spectrum, improving overall prediction precision. Furthermore, the system's ease of integration and adaptability set a new benchmark for emotional detection tools in virtual educational settings.

II. LITERATURE REVIEW

The integration of emotional detection in the educational field has grown significantly in recent years due to the importance of understanding students' emotional states and their impact on the learning process. Various studies have shown that emotions play a crucial role in motivation, academic performance, and active participation in the classroom [18]. This phenomenon has led to developing systems that seek to identify, analyze, and react to students' emotions using facial recognition, voice analysis, and biometric sensors.

Affectiva and Emotient [10], [19] are two of the best-known commercial solutions that have applied facial and voice recognition to detect emotions. Affectiva, for example, uses artificial intelligence (AI) to analyze individuals' facial expressions and tone of voice, offering solutions in various areas, including education. Several studies, such as the one conducted by Kulke et al. [20], have shown that the precision of emotions detected by Affectiva can reach up to 85%, consistent with Affectiva's observations in detecting emotions such as happiness, surprise, and sadness. However, the integration of Affectiva in educational environments has been limited, mainly due to the need for a customized technical adaptation in educational platforms such as Moodle or Microsoft Teams. This is a significant challenge, as academic institutions require solutions that are easy to implement and do not depend on complex technical support.

On the other hand, Emotient, before its acquisition by Apple, used facial analysis technology to measure emotions and was applied in areas such as health and safety. Although this solution has also shown high precision, its application in educational environments is less flexible due to its limitations in adaptability and scalability. Schmitz-Hübsch et al. [21] highlight that solutions such as Emotient are useful in controlled environments but present difficulties in large-scale applications, especially in large virtual classrooms.

Meanwhile, other works have focused on developing more accessible emotional detection systems for educational platforms. In this sense, the work of Ortega-Ochoa et al. [22] on using AI tools in education highlights how real-time emotional detection systems can identify emotions and provide instant feedback to students and teachers. These emotional feedback systems allow teachers to adjust their methodology or intervention based on students' emotional states. Compared to Affectiva and Emotient, the advantage of these systems is their ease of integration into existing educational platforms such as Teams or Zoom, making them easy to adopt and use immediately.

In our proposal, we have considered the advantages of these commercial solutions. Still, we have focused on optimizing educational integration in educational environments, offering a more accessible and scalable solution. Our tool stands out in its precision in emotion detection (with a performance comparable to that of Affectiva) but with a clear advantage in terms of ease of implementation and scalability for large classrooms. As indicated by Zhao et al. [23] adaptability is essential to ensuring that the emotional detection system is effective in small samples and can be successfully applied on a large scale in various educational environments.

III. MATERIALS AND METHODS

A. WORK ENVIRONMENT AND TECHNOLOGICAL TOOLS

The emotional detection system is integrated within Microsoft Teams, using the Microsoft Graph API to process real-time data such as interactions and audio and video signals. TensorFlow and PyTorch are used for voice analysis, with RNNs and LSTM models to capture emotions through vocal pitch and rhythm, using datasets such as RAVDESS [24], [25]. In facial expression recognition, OpenCV is employed for face detection, while Dlib extracts key landmarks, ensuring precise localization of facial components. These extracted features are then processed by CNNs specialized in facial expression analysis. The fusion of multimodal voice and face data uses specific neural networks to improve precision [26]. The interface is managed through the Microsoft Bot Framework and Azure Cognitive Services, providing real-time emotional feedback. The technical infrastructure includes high-quality cameras, microphones, and servers with enough power to process the deep learning models efficiently and with low latency, ensuring fluid interaction during virtual classes.

1) EDUCATIONAL PLATFORM: MICROSOFT TEAMS

Microsoft Teams has been selected as the core educational platform for implementing the emotional detection application due to its wide adoption in educational environments and its ability to integrate external applications using the Microsoft Graph APIs. Microsoft Teams facilitates online collaboration through meetings, chat, and content management in an accessible environment [16]. The developed application will be directly integrated into Teams using Microsoft Teams App Studio, a tool for creating customized applications for this environment.

The integration will be done by taking advantage of Teams' capabilities for real-time interaction during educational sessions. The emotional detection system will be integrated into Teams sessions to capture students' voice signals and facial expressions using the microphone and webcam. The data collected during interactions in virtual classes will be processed directly through the application within Teams, allowing immediate feedback on the student's emotional state. The emotions will be analyzed in real-time, and the results will be displayed to the teacher within the Teams interface through notifications or an interactive dashboard, providing relevant information about the student's emotional well-being.

Through the Microsoft Graph API, the application's interactions with Teams resources, such as video calls, participant data, and audio and video collection during online sessions, will be managed. In addition, the application can be integrated with Teams Bots, which will allow for automated interaction with students, making recommendations or alerting the teacher about possible emotional indicators of students during classes [27].

2) LANGUAGES AND FRAMEWORKS

For the application's development, Python and its specialized libraries are used, given their robustness and flexibility in processing voice signals and analyzing facial images, which are fundamental for the emotional detection system. Python has become the main language in research and development applications involving AI and multimodal data processing.

Voice signal processing uses libraries such as Librosa for acoustic feature extraction. Librosa is a Python library that provides tools for the extraction of audio features, such as fundamental frequency (F0), mel-spectrograms, cepstral coefficients at Mel frequencies (MFCC), and pitch [28].

These acoustic parameters are essential for detecting emotions through variations in the voice's pitch, rhythm, and volume. In this work's context, the primary function of this library is to extract the relevant features from voice data captured in real time during interactions in Microsoft Teams.

Facial expression analysis will be performed using the OpenCV library, which is fundamental in computer vision tasks. OpenCV allows real-time image capture and manipulation from the student's camera [29]. Detecting vital facial points allows facial expressions to be analyzed to infer emotions. Using a pre-trained model such as Haar Cascades, OpenCV allows for identifying the face in an image or video, facilitating the detection of specific facial features such as eyes, mouth, and eyebrows, which are key indicators of emotions.

Deep neural networks (DNNs) implemented with TensorFlow or PyTorch will detect and classify emotions from features extracted from voice and facial expressions. TensorFlow is a widely used machine learning framework for creating neural networks and other AI models, allowing for the efficient implementation of DNN models for emotional classification. TensorFlow Keras will be used to build emotion classification models from features extracted from speech signals and facial images [30].

PyTorch is another viable option for building neural networks due to its flexibility and ease of implementation, especially in research projects where model customization is essential. PyTorch will be used to create more advanced emotion classification models that integrate features from speech signals and facial expressions using RNNs or LSTMs, which deal with data sequences such as those obtained in speech processing.

Multimodal fusion techniques will be employed to improve the precision of emotional detection, using neural networks that combine the features extracted from the two types of data (voice and facial expressions). Early fusion models, where the features are combined before passing through a classification layer, or late fusion models, where the individual predictions from each modality are combined at the end to generate a joint prediction, can be applied. The goal is for the system to integrate voice data and facial expressions to improve the precision of emotional classification.

3) TECHNICAL INFRASTRUCTURE

The infrastructure required to run the tool is designed to ensure efficient real-time processing of voice signals and facial images. In terms of hardware, the system does not require specialized equipment, as standard devices used by students for virtual classes, such as webcams and microphones, will be leveraged. These devices will be sufficient to capture the necessary signals, provided they are used with appropriate resolutions and guaranteed minimum quality of 720p for the cameras and a sampling frequency of at least 44.1 kHz for the microphones. However, in scenarios with older hardware, such as low-resolution cameras or low-fidelity microphones, the system integrates pre-processing algorithms to compensate for quality degradation. Noise reduction techniques for audio and resolution upscaling for video streams are employed to enhance signal clarity. These measures ensure that emotional detection remains accurate even when using less advanced devices.

Real-time data processing is performed on a central processing server hosted in the cloud to ensure scalability and accessibility. This server uses instances with dedicated GPUs to execute neural network models, essential to handling the intensive computational loads associated with the real-time inference of DNNs. Platforms like Google Cloud or Microsoft Azure provide robust and accessible infrastructure [31]. The architecture also includes an edge processing component for initial data filtering and compression on the user's device. This minimizes latency and reduces the data transmitted to the cloud, ensuring real-time performance even in environments with limited bandwidth. To address poor internet connectivity, the edge processing component includes a local caching mechanism, temporarily storing and processing data on the device before transmitting it to the cloud. This feature ensures the system remains functional in resource-constrained settings, providing a consistent user experience regardless of internet quality.

Integration with Microsoft Teams is achieved through its publicly available APIs and SDKs, allowing the system to retrieve audio and video streams from live sessions seamlessly. The system processes these streams in real-time using event-driven architectures that efficiently handle the continuous data flow. The integration ensures that emotional data is extracted and analyzed without disrupting the ongoing class session, offering teachers insights into student engagement and emotional states in real time. Extensive testing has been conducted to ensure compatibility with various configurations of Teams, demonstrating robustness across different institutional setups.

Furthermore, the system benefits from Docker containers to ensure the application runs consistently across different environments, making it easy to deploy and maintain without relying on specific hardware configurations. These containers allow the emotional detection application to run efficiently and be scalable on any cloud infrastructure, regardless of the physical configuration of the server. Additionally, the containers are optimized with hardware acceleration frameworks such as NVIDIA CUDA and TensorRT to maximize performance during inference, enabling efficient use of available resources on GPU-equipped servers. To support scalability and compatibility with diverse hardware, the system dynamically adjusts processing pipelines based on the detected capabilities of the user's device, ensuring optimal performance even with hardware variability. The flexibility of this design ensures that institutions with varying levels of technological infrastructure can adopt the system without additional investments in hardware or connectivity.

B. DATA CAPTURE PROCESS

1) EXPERIMENTAL SETUP AND DATA SYNCHRONIZATION

The experimental setup was designed to ensure the systematic collection and evaluation of multimodal emotional data in a real-world educational context. The study involved 150 participants from a university-level virtual learning environment, including 120 students and 30 educators. Data collection spanned over six weeks, during which 45 online sessions, each lasting 60 minutes, were conducted. These sessions were carried out via Microsoft Teams to simulate actual class interactions and ensure ecological validity. All participants provided informed consent before data collection, adhering to ethical guidelines.

Facial images and voice signals were captured synchronously using standard webcams and microphones integrated into the participants' devices. Each video frame (captured at 30 frames per second) was paired with audio segments of the corresponding time frame to ensure temporal alignment between the modalities. For voice signals, audio data was segmented into overlapping windows of 25 milliseconds with a stride of 10 milliseconds, capturing both short-term and transitional vocal features. Timestamps embedded in the video and audio streams facilitated precise alignment during preprocessing, ensuring that features from both modalities corresponded to the same interaction period.

Feature extraction processes were synchronized to handle the varying sampling rates and voice and image data dimensionalities. Facial features, such as eyebrow movement and lip curvature, were extracted from each video frame. In contrast, prosodic and spectral audio features, including MFCCs and pitch variations, were computed for the corresponding audio window. Dimensional normalization techniques ensured that both modalities shared a consistent representation scale, enhancing compatibility during the multimodal fusion process.

The model's predictions were validated using expert assessments as a reference. Psychological experts observed the recorded sessions and manually annotated emotions using a predefined rubric covering seven emotional categories: happiness, sadness, stress, calm, frustration, surprise, and interest. The annotations were compared with the system's real-time predictions, and discrepancies were analyzed to refine the alignment and classification processes. The alignment process ensured a robust evaluation by capturing dynamic emotional changes across modalities and timeframes.

Figure 1 illustrates the experimental setup and data synchronization pipeline, highlighting the steps involved in capturing, aligning, and processing multimodal emotional data. This architecture ensures the feasibility and scalability of the proposed approach while maintaining the rigor required for real-time applications.

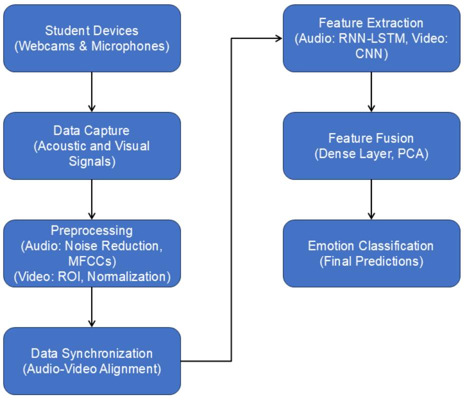
Student Devices (Webcams & Microphones)
Data Capture (Acoustic and Visual Signals)
Preprocessing (Audio: Noise Reduction, MFCCs) (Video: ROI, Normalization)
Data Synchronization (Audio-Video Alignment)
Feature Extraction (Audio: RNN-LSTM, Video: CNN)
Feature Fusion (Dense Layer, PCA)
Emotion Classification (Final Predictions)

FIGURE 1. Experimental setup and data synchronization pipeline for multimodal emotional detection.

2) VOICE DATA

Voice data is captured through the microphone built into the devices used by students in online class interactions via platforms such as Microsoft Teams. The captured audio signal is processed using advanced acoustic signal processing techniques to extract relevant features that allow the speaker's emotions to be identified. The extracted features include prosody parameters such as pitch, rhythm, and intensity and spectral features such as MFCCs. These parameters are considered fundamental for emotional detection since variations in the pitch and rhythm of the voice are directly related to emotions such as joy, sadness, frustration, or stress.

The processing of these data is carried out using the Librosa package in Python, which allows the extraction of acoustic features such as MFCCs, Chroma features, and spectral contrast, representing key characteristics of the human voice [32]. These extracted feature vectors have a fixed dimensionality of 40 MFCC coefficients per frame, computed over 25 ms windows with a stride of 10 ms, ensuring fine-grained temporal resolution.

The extracted acoustic features serve as inputs for a recurrent neural network (RNN) with LSTM layers designed to capture temporal dependencies in emotional variations by learning sequential patterns in speech. This architecture consists of two bidirectional LSTM layers with 128 units each, allowing the model to capture both forward and backward dependencies in the speech signal. A dropout rate 0.3 is applied after each LSTM layer to prevent overfitting. The final dense layer maps the extracted features to a softmax classification output, predicting one of seven emotional states (happiness, sadness, stress, calm, frustration, surprise, and interest).

To ensure robustness in the extracted features, preprocessing steps are applied to the raw audio data, including:

- Noise Reduction: A spectral subtraction algorithm removes background noise.
- Signal Amplitude Normalization: Ensures consistency across different recording environments.
- Voice Activity Detection (VAD): Eliminates silent regions to focus on meaningful speech.
- Audio Segmentation: Each speech sample is segmented into fixed-length frames (1.5 seconds) aligned

with corresponding video frames, ensuring precise synchronization.

This structured approach ensures that the extracted speech features are effectively processed by the LSTM model, allowing for accurate emotion classification while maintaining robustness across diverse recording conditions.

3) FACIAL EXPRESSION DATA

Facial expression capture uses the student device's webcam, which transmits images or videos during class interactions. The system applies computer vision techniques to these images to identify key facial features that indicate emotional states. To detect faces, OpenCV is used, employing Haar cascades and DNN–based models to locate the face within the image. Once detected, Dlib is applied to extract key facial landmarks such as the position of the eyes, mouth, and eyebrows.

CNN models are ideal for image analysis because they can identify spatial patterns within facial images. These convolutional networks process facial images, identifying facial features such as eyes, mouth, and eyebrows that indicate specific emotions.

To detect faces, OpenCV is used, employing Haar cascades and DNN–based models to locate the face within the image. Once detected, Dlib extracts key facial landmarks, such as the position of the eyes, mouth, and eyebrows, which are then processed to generate numerical descriptors of facial structure. These extracted features serve as input for CNNs trained explicitly for emotion recognition. The CNN models were trained and fine-tuned for this implementation using AffectNet and RAVDESS datasets.

AffectNet was used to train CNN on various facial expressions, including happiness, sadness, anger, surprise, fear, and disgust. RAVDESS, on the other hand, was incorporated to enhance robustness in dynamic emotional expressions. It includes facial expressions synchronized with speech, allowing better adaptation to real-world variations in facial gestures.

This dataset was employed to train the CNN models from scratch and fine-tune pre-trained architectures. The models were further optimized for real-time classification by performing additional training using a dataset specifically curated for the educational context, capturing variations in frustration, stress, and student engagement.

The raw facial image data undergoes preprocessing to improve the quality and reliability of the extracted features. This involves resizing the images to a standardized resolution, converting them to grayscale to reduce computational complexity, and applying histogram equalization to normalize lighting conditions. Additionally, facial alignment is performed using landmark-based transformation techniques to minimize variations due to head pose differences. These preprocessing steps ensure that the CNN models focus on essential spatial patterns while mitigating variability due to environmental factors such as lighting or camera angles.

4) IMAGE PROCESSING TECHNIQUES

Facial image processing involves a series of steps in which the face is segmented, and the relevant facial features for emotional analysis are extracted. The first step is facial detection, using Haar Cascades in OpenCV, a technique that allows the position of the face in images to be localized. A deep learning-based face detector from OpenCV's DNN module is also employed to enhance robustness in varied lighting conditions and poses. Once the face is detected, the region of interest (ROI) is extracted using the facial landmarks detected by Dlib. The bounding box defining the ROI is dynamically adjusted based on the Euclidean distances between the eyes and mouth, ensuring a standardized and consistent input size [33]. The ROI extraction process applies bounding box adjustments and geometric transformations to provide precision localization of facial components.

After segmentation, Dlib is utilized to extract 68 facial landmarks, which provide precise spatial information about key facial regions such as the contour of the lips, the position of the eyebrows, and the corners of the eyes. Dlib was chosen over MediaPipe due to its higher accuracy in structured facial analysis and stability in video sequences. While MediaPipe is optimized for lightweight applications, Dlib provides a more detailed and consistent representation of facial geometry across different frames. These landmarks serve as the input for further feature extraction processes. Dlib does not classify emotions but ensures consistent tracking of facial key points across frames, allowing a structured representation of facial geometry.

A facial feature extraction algorithm is then applied to obtain the details that will be used for emotional classification. The extracted features include Euclidean distances between key facial landmarks (e.g., eye-to-eye distance, mouth width), geometric analysis of lip curvature, and relative eyebrow positions. These values are structured into numerical feature vectors representing spatial relationships among facial components. These numerical representations are structured into feature vectors that serve as input to a CNN specialized in facial expression analysis. The extracted features are transformed into structured 1-D feature vectors to optimize classification, normalized, and concatenated with the processed ROI image. This enables the CNN to analyze spatial and numerical patterns simultaneously, improving emotion classification precision.

In addition to conventional facial expression analysis, the system incorporates microexpression detection to capture subtle and involuntary emotional cues. These microexpressions, which last between 40 and 200 milliseconds, provide deeper insights into emotions such as frustration, stress, and surprise, which may not be as apparent in standard facial expressions.

The detection of microexpressions is achieved through a combination of Optical Flow and CNN-based models trained on CASME II and SMIC. Optical Flow captures fine-grained motion variations in facial skin, identifying muscle tension shifts in the forehead, around the eyes, and in the perioral area. The detected motion patterns are then processed by CNN models, specifically fine-tuned datasets, ensuring high sensitivity to short-lived facial movements.

The extracted microexpression features are represented as motion flow maps and fed into the CNN for classification into emotional categories such as frustration, surprise, stress, or distrust. Including these features enhances the model's ability to detect subtle, involuntary emotions, contributing to a more precise assessment of student engagement and affective states.

Voice and facial expression data are synchronized using temporal alignment techniques to ensure the features from both modalities correspond to the same time frames. This synchronization is achieved by matching video frames with audio frames through timestamp interpolation, ensuring each facial expression aligns precisely with its corresponding vocal segment. As depicted in Figure 2, this process ensures a structured integration where both modalities contribute simultaneously to the final emotional classification rather than being processed independently. To reinforce the fusion process, extracted features from the audio and facial analysis are combined in a shared representation before classification, preventing misalignment between audio, microexpressions, and microexpressions.

The emotional detection system's data capture process relies on collecting acoustic signals through microphones and facial images through cameras. These data are processed using advanced signal processing and computer vision techniques, employing specialized neural networks to extract emotional features and classify emotions in real time. By integrating synchronized voice data, macro-expressions, and microexpressions, the system achieves a comprehensive multimodal feature set that significantly improves the precision and robustness of emotion recognition in educational environments.

C. MULTIMODAL FUSION APPROACH

Integrating facial image features and voice signal characteristics is a cornerstone of the proposed emotional detection system. This provides complementary information that enhances the precision and robustness of emotional detection. This section elaborates on the fusion process's conceptual framework, methodologies, and technical implementation, unifying modalities into a cohesive representation.

The fusion process begins with feature extraction from each modality, ensuring that relevant emotional cues are captured from facial expressions and voice signals.

- Facial Features: Extracted using a CNN trained on AffectNet, which detects facial landmarks and microexpressions, which are critical for identifying subtle and involuntary emotional cues. Within 40 to 200 milliseconds, these microexpressions provide insights into emotions such as frustration, stress, and surprise, often not captured in standard facial expression analysis. Their detection is based on Optical Flow analysis combined with CNNs trained on CASME II and SMIC, ensuring robust recognition of rapid facial changes.
- Voice Features: Extracted through an RNN with LSTM layers, leveraging temporal dependencies in acoustic signals. The audio features include MFCCs, pitch, and energy metrics, which are crucial for capturing vocal emotionality. The RNN consists of two bidirectional LSTM layers with 128 units each, allowing the model to track variations over time. Dropout regularization (0.3) is applied to prevent overfitting.
- Dimensionality Reduction: Principal Component Analysis (PCA) is applied to facial image features to ensure compatibility between modalities. PCA reduces computational complexity while retaining the most significant variation in the data, keeping 95% of the explained variance with approximately 50 principal components.
- Temporal Alignment: Audio and video signals are synchronized to ensure both modalities contribute relevant information per frame. Timestamp interpolation precisely aligns extracted voice and facial features in corresponding time frames.

The combined feature representation is constructed using early fusion, where extracted features from both modalities are concatenated into a unified vector, preserving key emotional cues before classification.

Mathematically, this can be represented as:

$$\mathbf{F}_{\text{fusion}} = \sigma(\mathbf{W}_1[\mathbf{F}_{\text{audio}}, \mathbf{F}_{\text{image}}] + \mathbf{b}_1) \tag{1}$$

Here, $\mathbf{F}_{\text{audio}} \in \mathbb{R}^{d_1}$ represents the audio feature vector, including MFCCs, pitch, and energy metrics, $\mathbf{F}_{\text{image}} \in \mathbb{R}^{d_2}$ represents the facial feature vector, composed of landmark positions, microexpression encodings, and PCA-reduced features. $\mathbf{F}_{\text{fusion}} \in \mathbb{R}^{d_1+d_2}$ is the final fused representation, which is passed through a fully connected layer to learn modality-specific interactions.

The fused vector is then fed into a fully connected neural network (DNN) for classification. The network consists of:

- ReLU functions activate two hidden layers of 256 and 128 neurons.
- Dropout (0.3) after each hidden layer to prevent overfitting.

A final softmax layer for classification into seven emotional categories:

$$y = \text{Softmax}(\mathbf{W}_2 \cdot \mathbf{F}_{\text{fusion}} + \mathbf{b}_2) \tag{2}$$

Transfer learning is employed to enhance model generalization. The CNN and LSTM models are initialized with pre-trained weights from AffectNet, RAVDESS, and CASME II. AffectNet is used for macro-expression training in static images. RAVDESS enhances speech-associated emotional recognition. CASME II and SMIC fine-tune microexpression recognition, ensuring the model captures involuntary emotional cues. The models are fine-tuned using an internally curated dataset from educational environments,

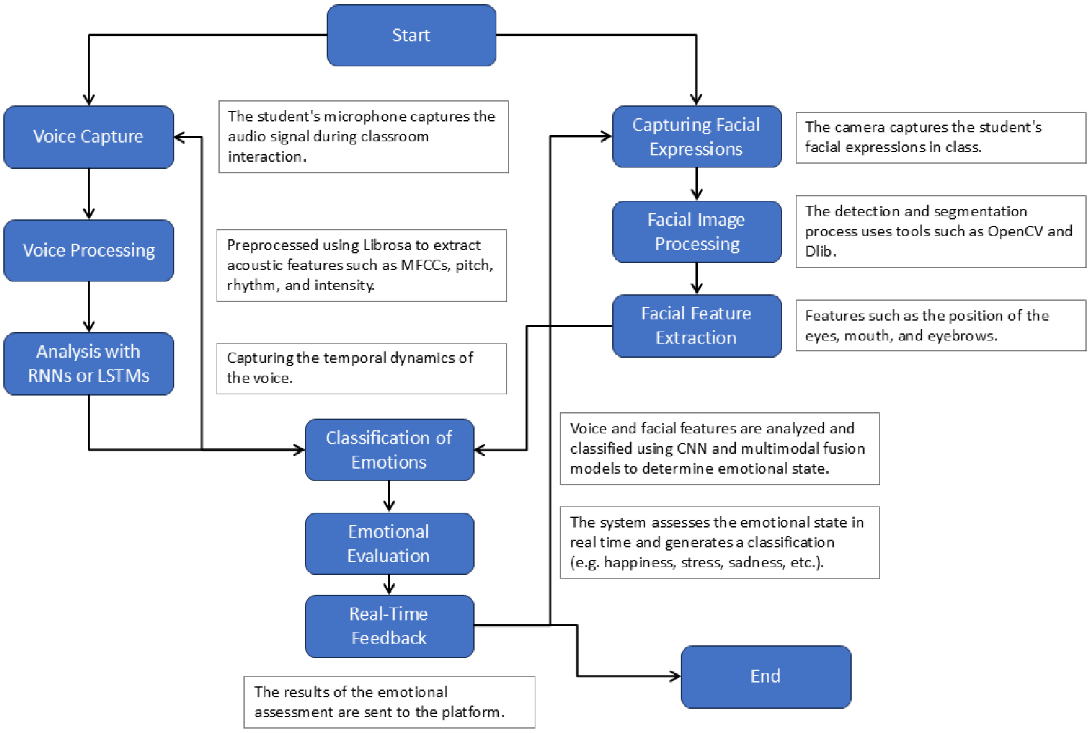
Start
Voice Capture
The student's microphone captures the audio signal during classroom interaction.
Voice Processing
Preprocessed using Librosa to extract acoustic features such as MFCCs, pitch, rhythm, and intensity.
Analysis with RNNs or LSTMs
Capturing the temporal dynamics of the voice.
Capturing Facial Expressions
The camera captures the student's facial expressions in class.
Facial Image Processing
The detection and segmentation process uses tools such as OpenCV and Dlib.
Facial Feature Extraction
Features such as the position of the eyes, mouth, and eyebrows.
Classification of Emotions
Voice and facial features are analyzed and classified using CNN and multimodal fusion models to determine emotional state.
Emotional Evaluation
The system assesses the emotional state in real time and generates a classification (e.g. happiness, stress, sadness, etc.).
Real-Time Feedback
The results of the emotional assessment are sent to the platform.
End

FIGURE 2. Data capture and processing flow for emotional detection.

providing better adaptation to student emotions such as frustration, stress, and engagement.

Additionally, late fusion methods were evaluated as a comparative approach, where the independent predictions from voice and facial expression models were combined using a weighted average:

$$P_{\text{fusion}} = \alpha P_{\text{voice}} + (1-\alpha) P_{\text{facial}} \tag{3}$$

Here, α is a weighting parameter that adjusts the relative importance of each modality. In microexpression-based features were considered an additional factor. Still, they were found to be less effective in a late fusion framework due to their short duration, which requires precise temporal synchronization. Thus, early fusion was selected as the primary method for integrating all modalities.

The multimodal fusion architecture handles the variability and limitations inherent in individual modalities. For instance, if facial expressions are occluded or ambiguous, audio features provide complementary cues, and vice versa. Microexpression analysis further enhances the system's sensitivity to involuntary emotional cues, improving emotional classification's overall depth and accuracy. This integration captures nuanced patterns undetectable with unimodal approaches, improving the overall precision and robustness of the emotional detection system. Furthermore, the design ensures scalability and efficiency, making it suitable for real-time applications in educational settings where synchronized data streams must be processed with minimal latency.

D. EMOTIONAL DETECTION ALGORITHMS

1) EMOTIONAL CLASSIFICATION MODELS

The emotion detection system employs multiple classification algorithms to process multimodal data from voice signals and facial expressions. The selected models address different aspects of the classification task, ensuring a robust and precise analysis of emotional states. The system integrates Support Vector Machines (SVMs), CNNs, and RNNs with LSTM layers, each optimized for the type of data it processes.

SVMs classify structured numerical features derived from voice and facial expressions. Acoustic features, such as MFCCs, pitch variations, and speech intensity, are converted into high-dimensional feature vectors. Similarly, geometric facial features, including distances between facial landmarks and variations in lip curvature, are structured for classification [36]. The SVM model learns an optimal hyperplane that separates different emotional classes by maximizing the decision margin in the feature space. This classification is formulated mathematically as follows:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{such that} \quad y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1, \forall i \tag{4}$$

where $\mathbf{x_i}$ is a feature vector (e.g., voice or facial features), y_i is the associated emotional label, $\mathbf{w}$ is the weight vector, and b is the bias.

For facial expression recognition, CNNs are employed to process images captured during class interactions [37]. These models automatically extract spatial features from facial images, identifying key patterns in expressions such as eyebrow movement, lip shape, and eye openness. The architecture consists of an input layer that normalizes grayscale or RGB images to a resolution of 48 × 48 pixels, followed by multiple convolutional layers with 3 × 3 filters and ReLU activation. Batch normalization layers stabilize training while pooling layers reduce dimensionality. Fully connected layers with 128 and 64 neurons refine the extracted features before passing them through a softmax classifier that assigns the probability of each emotion category. The convolution operation within each layer is defined as:

$$\mathbf{y_i} = \sigma(\mathbf{W}\mathbf{x_i} + b) \tag{5}$$

where $\mathbf{W}$ is the convolution filter, $\mathbf{x_i}$ is the input (in this case, a facial image), σ is the activation function, and b is the bias.

RNNs and LSTMs are employed to capture temporal dependencies in voice data. These models analyze sequential acoustic features, allowing the detection of dynamic changes in pitch, energy, and rhythm that correspond to emotional states. The system utilizes a bidirectional LSTM architecture with two layers containing 128 hidden units with ReLU activation. Dropout regularization at a rate of 0.3 is applied to prevent overfitting. The following equation governs the LSTM unit's operation:

$$\mathbf{h_t} = \sigma(\mathbf{W} \cdot \mathbf{h_{t-1}} + \mathbf{U} \cdot \mathbf{x_t} + \mathbf{b}) \tag{6}$$

where $\mathbf{h_t}$ is the hidden state at time t, $\mathbf{h_{t-1}}$ is the previous hidden state, $\mathbf{x_t}$ is the input at time t, $\mathbf{W}$ and $\mathbf{U}$ are weight matrices, and σ is an activation function (usually sigmoid or tanh). These LSTM models are trained using the RAVDESS dataset, which provides labeled speech recordings containing emotional variations. Speech samples collected from real educational settings are fine-tuned to enhance adaptability to the target domain.

The integration of these models follows a structured multimodal approach, where CNNs process facial expression features, LSTMs analyze temporal variations in voice data, and SVMs serve as a comparative model for structured numerical feature classification. Feature fusion occurs in the later stages of the pipeline, where embeddings from CNNs and LSTMs are concatenated into a unified representation. Hyperparameter optimization, including tuning learning rates (ranging between 0.0001 and 0.001), adjusting batch sizes (32, 64, 128), and selecting optimal kernel sizes for convolutional layers (3 × 3 vs. 5 × 5), ensures the models generalize effectively across different datasets. Cross-validation with $k = 5$ is applied to assess model robustness, and data augmentation techniques such as noise addition, pitch shifting for audio, and geometric transformations for facial images enhance training diversity.

By leveraging these models in a complementary manner, the system achieves a refined and adaptive emotional detection capability suited for real-time applications in educational environments. The combined methodology ensures a precise and scalable approach to monitoring students' emotional engagement and response during virtual learning sessions.

2) MODEL TRAINING

The emotional classification models are trained using datasets containing labeled examples of emotions expressed in voice and facial expressions. The training data includes multiple voice samples from students in different emotional contexts, such as happiness, sadness, stress, and frustration, as well as facial images depicting these emotions.

RAVDESS, a dataset containing emotional audio recordings in English, is used to train the voice models. This dataset provides a comprehensive range of vocal emotions, including happiness, sadness, fear, and surprise. The training process for LSTM models includes feature extraction steps where MFCCs, pitch variations, and energy levels are computed per frame. The extracted features are structured into time-series sequences and fed into an LSTM model with three layers, each containing 128 hidden units, followed by a dense layer for classification. These models are particularly effective for temporal sequence analysis, capturing long-term dependencies within the audio data to detect emotional patterns.

AffectNet, a widely used dataset for facial emotion recognition, is utilized for facial expressions. It contains over a million facial images labeled with different emotions, allowing CNN models to be trained to classify facial expressions into different emotional categories. The CNN architecture consists of five convolutional layers with ReLU activation, each followed by max-pooling layers to reduce spatial dimensions. The final feature maps are flattened and passed through two fully connected layers before classification. The system integrates additional labeled data from educational contexts where these emotions are prevalent to address complex emotions such as frustration and stress. Fine-tuning is performed using a subset of AffectNet combined with an internally curated dataset, ensuring the model is specifically adapted to the nuances of educational environments. Datasets such as DAiSEE, designed explicitly for affective states in academic environments, are incorporated to enrich the training process. Moreover, creating a custom dataset based on real-world educational interactions is proposed. This dataset includes voice recordings and facial expressions captured during classroom activities, tailored to reflect the nuanced dynamics of frustration and stress in learning environments.

PCA reduces the dimensionality of feature representations extracted from CNNs and LSTMs. The number of principal components is determined by retaining 95% of the variance, ensuring that only the most relevant features contribute to the classification model.

Augmentation techniques such as pitch shifting, noise addition for audio, and geometric transformations for facial images are applied to diversify the training data and improve robustness. For audio data, pitch is shifted within a range of ±2 semitones, and white noise is added with a signal-to-noise ratio (SNR) of 20 dB. In facial images, geometric transformations include random rotation (±10 degrees) and horizontal flipping (50% probability). Transfer learning is employed, initializing the model with weights pre-trained on large-scale datasets like AffectNet and fine-tuning it with domain-specific data to enhance its capacity to identify nuanced emotional expressions in an educational setting.

Hyperparameter tuning is conducted using a grid search approach. Learning rates are tested from 0.0001 to 0.001, batch sizes of 32, 64, and 128 are evaluated, and dropout rates between 0.2 and 0.5 are optimized to prevent overfitting.

a: INTEGRATION OF MULTIMODAL DATA

A multimodal training pipeline enhances the detection of emotions, particularly complex categories like stress and frustration. Features extracted from voice data and facial expressions are concatenated into a unified vector using a dense fusion layer. This layer captures modality-specific interactions and generates a combined representation of the emotional features.

b: OPTIMIZATION AND REGULARIZATION

Models are trained using optimization algorithms such as Adam or Stochastic Gradient Descent (SGD), which adjust the neural network weights to minimize the loss function. Cross-entropy is used as the primary loss function for classification tasks. Class imbalance is addressed using focal loss, which reduces the relative impact of easily classified samples, ensuring better performance on underrepresented emotions. In contrast, weighted loss components are introduced to balance the contributions of rare emotional categories, such as frustration and stress. Regularization techniques, including Dropout and L2 regularization, are applied to mitigate overfitting and ensure that the models generalize well to new data.

c: CURRICULUM AND MULTI-TASK LEARNING

The training process employs curriculum learning, which gradually increases the difficulty of training examples over successive epochs to improve the detection of subtle and complex emotions. Initially, the model is trained on highly distinguishable emotions (e.g., happiness vs. sadness) before progressing to more ambiguous categories (e.g., frustration vs. stress). This approach lets the model learn simpler patterns, building a foundation for identifying more intricate emotional cues. Multi-task learning is also integrated, enabling the model to jointly optimize for emotion classification and intensity estimation tasks. This approach improves the model's ability to recognize overlapping features, such as the shared characteristics between stress and frustration.

d: TRAINING AND VALIDATION

The model training pipeline uses cross-validation techniques to ensure robustness and prevent overfitting. The data is split into multiple folds, with the model trained on a subset of the data and validated on the remaining fold. A five-fold cross-validation strategy is applied, where each fold serves as a test set once while the remaining data is used for training. Performance metrics, including precision, recall, and F1-score, are monitored during training to evaluate the model's effectiveness. Confidence intervals for these metrics are computed using bootstrapping, ensuring statistical reliability in the reported results.

3) MODEL EVALUATION

Model precision is evaluated using standard classification metrics, which are applied to measure the performance of models on test data that was not used during training. The most relevant evaluation metrics include:

Precision: Measures the proportion of correct predictions about the total predictions made:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

where TP is the number of true positives, and FP is the number of false positives.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

where FN is the number of false negatives.

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

For complex emotions such as frustration and stress, the system evaluates performance using confusion matrices to visualize errors and misclassifications for these classes. The confusion matrix provides a detailed breakdown of the number of true positives, false positives, true negatives, and false negatives for each emotional category, enabling targeted analysis of system limitations.

Class imbalance is mitigated through weighted loss functions to ensure that rare emotional categories, such as frustration and stress, are accurately represented in predictions. Furthermore, multi-task learning is explored, where the model simultaneously optimizes for multiple related tasks, such as emotion classification and intensity estimations, improving sensitivity to subtle patterns in complex emotions.

To strengthen the evaluation's statistical robustness, confidence intervals for each metric (e.g., precision, recall, and F1-score) are calculated using bootstrapping techniques. This involves repeatedly resampling the test data and computing the metrics across multiple iterations to estimate variability and provide 95% confidence intervals for each metric. These intervals quantify the uncertainty of the performance metrics, ensuring that the results are statistically reliable.

Cross Validation: This method ensures the model's robustness by splitting the data into several subsets (folds), training and evaluating the model on each fold in a rotating manner. This allows for a more reliable assessment of the model's performance, avoiding overfitting and ensuring that the model generalizes well to new data. Additionally, sensitivity analysis is conducted to evaluate the impact of variations in hardware quality (e.g., reduced audio fidelity or lower video resolution) on model performance, ensuring robustness in real-world applications. This analysis informs hardware recommendations for achieving optimal performance under different deployment scenarios.

Finally, statistical significance tests, such as paired t-tests or Wilcoxon signed-rank tests, are applied when comparing the model's performance against baseline systems or alternative configurations. These tests assess whether observed differences in metrics, such as F1-scores for frustration and stress, are statistically significant, reinforcing the reliability of the findings.

E. APPLICATION DEVELOPMENT IN MICROSOFT TEAMS

1) INTERFACE DESIGN

The app's user interface (UI) within Microsoft Teams is designed to be intuitive and easily accessible during online interactions. The app integrates directly into the Teams sidebar, allowing teachers and students to interact with the tool without leaving the platform. The central part of the interface is the options panel, where the teacher can access different functionalities, such as the visualization of emotional results and feedback tools. The Emotion Detector options allow the teacher to monitor the student's emotional state in real-time.

The results of the emotional detection are visually presented in real-time through an emotional state icon next to the student's name. This icon indicates the emotion detected (e.g., happiness, stress, sadness, etc.), allowing the teacher to adjust the interaction immediately. In addition, if the system detects a critical emotion, such as an elevated stress level, an alert is displayed in the interface, recommending possible actions the teacher could take to support the student, such as offering a break or adjusting the class content.

2) REAL-TIME INTERACTION

The system is designed to work interactively in real-time. The Emotion Detector app continuously receives data on student voices and facial expressions during class. Using the capabilities of the Microsoft Graph API, the app accesses the online session data, captures the audio and video signals, and processes them dynamically. The emotional detection model analyzes voice characteristics (such as pitch, rhythm, and intensity) and facial characteristics (such as eye, mouth, and eyebrow movement) to determine the student's emotional state.

The results of this analysis are processed and presented in the Teams interface in real-time, providing the teacher with immediate visual information about the student's emotional state. If the system detects a critical emotional state, such as stress or disconnection, an alert is triggered, displaying a recommendation to intervene [37]. The teacher can adjust the pedagogical approach or interact directly with the students to better understand their emotional state and offer appropriate support.

Figure 3 presents the application workflow, illustrating how voice and facial expression data are captured and processed in real-time and how emotional results are displayed on the platform. The figure also shows how alerts and suggestions are automatically generated to support the teacher in making decisions during class.

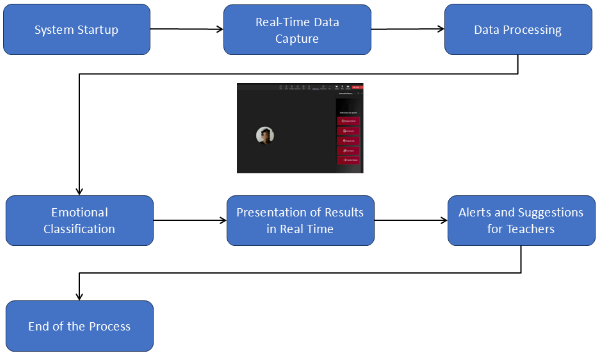

FIGURE 3. Workflow of the real-time emotional detection system.

Using this interface ensures that the teacher has access to crucial information about students' emotional well-being without disrupting the normal flow of the class. Thanks to this seamless integration within Microsoft Teams, the tool can facilitate a more personalized and adaptive learning experience, optimizing students' emotional environment while maintaining active engagement in the educational content.

This allows for continuous interaction between the system and students, facilitating the early identification of emotional issues and improving the teacher's ability to provide a more attentive and responsive educational environment to students' emotional needs.

F. SYSTEM EVALUATION AND VALIDATION

1) EVALUATION OF EMOTIONAL ACCURACY

System validation is performed by evaluating the accuracy with which the application predicts students' emotional state, comparing the model's predictions with the assessments of emotion experts, such as psychologists or educators. A pre-labeled dataset is used for this process, where experts have manually classified students' emotions. These experts review the emotional predictions generated by the system and assess their accuracy based on the match with human emotions observed in the educational context.

To measure the emotional accuracy of the system, the Pearson correlation coefficient (r) is used, which quantifies the linear relationship between the model's predictions and the experts' assessments. Mathematically, it is calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{10}$$

where y_i are the emotions rated by experts, $\hat{y}_i$ are the emotions predicted by the system, and $\bar{y}$ and $\bar{\hat{y}}$ are the means of the observed and expected emotions, respectively. An r value close to 1 indicates a high correlation and, therefore, higher emotional accuracy of the system.

In addition, the accuracy and precision of the model in classifying specific emotions, such as happiness, sadness, stress, etc., are evaluated. These metrics are calculated using standard definitions from classification theory, where accuracy is the percentage of correct predictions out of the total predictions, and precision refers to the ratio of true positives to optimistic predictions made by the model.

2) TESTING IN THE EDUCATIONAL ENVIRONMENT

Pilot tests are conducted in a controlled educational environment, using real students from diverse characteristics and backgrounds. The tests aim to validate the system's applicability and effectiveness in real-time teaching and learning situations. The selection of students for pilot tests is based on criteria such as:

- Emotional diversity: Students with different emotional characteristics are selected to ensure that the system can detect various emotional states in other contexts.
- Educational diversity: Students come from different disciplines and educational levels, ensuring the system is helpful in a diverse learning environment.
- Informed consent: All participating students must give informed consent to participate in the evaluation of the system and are assured that their data will be used only for research purposes.

The sample size is determined by a statistical power calculation, ensuring that the number of participants is sufficient to obtain meaningful results. Generally, a sample size of at least 30 students is set for each test group, allowing for sufficient data diversity to evaluate the system's performance.

The duration of the tests varies depending on the specific objectives, but they are typically conducted over 2 to 4 weeks. During this time, students interact with the system in various educational activities while teachers monitor alerts and suggestions generated by the system.

Specific evaluation metrics are used to evaluate the system's effectiveness, such as the accuracy of emotional predictions, the rate of student participation, and teacher satisfaction with the recommendations generated by the system. These metrics allow for analyzing how the system's emotional predictions affect pedagogical decisions and the dynamics in the virtual classroom.

3) ANALYSIS OF RESULTS

The system evaluation results are analyzed using several qualitative and quantitative approaches. At the end of the pilot tests, user feedback (both students and teachers) is collected through surveys and interviews. This feedback provides information on the tool's usability, effectiveness in improving the learning experience, and any difficulties participants perceive.

Quantitative metrics obtained during the pilot tests include:

- Emotional precision: Measured using the abovementioned metrics, such as precision, recall, and the confusion matrix. This evaluation quantifies the system's performance in classifying students' emotions.
- Teacher satisfaction: Using a Likert scale, teachers rate the usefulness of the system's alerts and recommendations. The average of the ratings is calculated to obtain an overall satisfaction score.
- Impact on student participation: The participation rate in interactive activities during classes (such as answering questions or discussions) is analyzed, correlating this data with emotional predictions. An increase in student engagement when a positive emotion, such as interest or curiosity, is detected would indicate that the system is helping to keep students engaged.

To interpret the results, accurate statistical analysis is conducted using hypothesis tests, such as the t-test or analysis of variance (ANOVA), to assess whether the observed differences in engagement and satisfaction metrics are significant [38] these analyses allow comparing test results with control groups or historical data from online interactions without using the emotional detection tool.

G. ETHICAL AND PRIVACY CONSIDERATIONS

1) PROTECTION OF PERSONAL DATA

Protecting the privacy of personal data is a critical priority in the development and implementation of this app. Since the emotional detection system collects and processes sensitive data, such as students' facial expressions and voice signals, strict measures must be taken to ensure that the collection and use of this data meet the highest privacy and security standards.

Firstly, all voice data and facial images collected by the app are anonymized to prevent the direct identification of students. This means that only emotional characteristics and emotional state predictions are stored instead of storing images or recordings directly associated with the student. Furthermore, it is ensured that facial images are used exclusively for emotion analysis and are not stored or shared with third parties.

The system complies with the European Union's General Data Protection Regulation (GDPR), which states that personal data must be processed lawfully, transparently, and with the explicit consent of the data subject. Furthermore, the system complies with local laws on privacy protection in education, ensuring that all data collection and use procedures comply with region-specific regulations.

Specific measures implemented to ensure data protection include:

- Data Encryption: All voice data and facial images are encrypted both in transit and at rest, using modern encryption standards such as AES-256 to ensure that unauthorized persons cannot intercept or access them [39].

- Secure Storage: Processed and stored data is stored on secure servers with restricted access, and periodic audits verify that security policies are adequately complied with.
- Data Destruction: Personal data is securely deleted once it has been processed and is no longer necessary for the purpose it was collected, thus ensuring that sensitive information is not stored for prolonged periods.

2) INFORMED CONSENT

Obtaining informed consent from students is an essential process that must be done clearly and transparently before collecting any emotional data. At the beginning of participation in the application, a consent form is presented, which explicitly details the nature of the data that will be collected (voice and facial expressions), the purpose of the collection (emotional analysis to enhance the educational experience), and the privacy measures that have been implemented to protect the data.

This form must be actively accepted by each student (or legal guardian in the case of underage students) before the data collection process begins. In addition, students can opt-out if they wish to, ensuring that the tool is not mandatory. Students who choose not to participate will not be excluded from the educational process and will be offered alternatives to continue participating in class activities without needing to use the emotional detection application.

Informed consent includes the following key points:

- Right to Withdraw: Students can withdraw their consent at any time without affecting the lawfulness of the data processing before withdrawal. If students decide not to continue participating, their data will be deleted immediately.
- Transparency: Students are ensured to have access to all relevant information about using the application, including the potential risks related to collecting facial and voice data and how this data will be used for emotional classification.
- Access to Data: Students can request information about the data collected and how it is being used at any time, as well as the possibility of correcting any errors in the processed data.

It should be noted that the system strictly limits the use of facial images to extract emotional features and does not use them for any other purpose. Facial identifications are not performed, and facial images are transformed into emotional features that do not allow the identification of students outside the context of the application.

3) ETHICAL IMPLICATIONS OF EMOTIONAL DETECTION

While emotional detection systems offer significant opportunities to enhance the educational experience, their implementation raises critical ethical concerns beyond privacy and consent. These include:

- Potential for Misuse of Emotional Data: Emotional data could be misinterpreted or used to enforce compliance rather than support learning. For example, detecting frustration or stress might inadvertently label a student as uncooperative or disengaged, creating unintended biases. To address this, the system explicitly limits its use to providing actionable insights for improving the learning experience and ensures it is not used for disciplinary actions or evaluations.
- Impact on Student Autonomy and Comfort: Students may feel pressured to participate due to the perceived authority of educational institutions, potentially compromising voluntary consent. To mitigate this, the system explicitly allows students to opt out without academic consequences and ensures equal access to alternative participation methods.
- Emotional Monitoring and Well-being: Continuous emotional monitoring may cause students discomfort or heightened self-awareness, impacting their natural behavior. To reduce undue stress, the system is designed to collect data only during specific educational activities and ensure that students are informed about when emotional monitoring occurs.
- Cultural and Individual Variability in Emotional Expression: Emotional detection systems must account for cultural and individual differences in emotional expression to avoid bias. The model incorporates training on diverse datasets to ensure inclusivity and fairness across cultural and individual contexts.

Addressing these ethical implications, the system aims to balance using emotional data to improve education while safeguarding students' rights, autonomy, and well-being.

IV. RESULTS

A. ASSESSMENT OF EMOTIONAL PRECISION

Evaluating emotional precision demonstrates the system's ability to classify emotions based on voice and facial data precision. The results highlight strong performance in detecting positive and neutral emotions, such as happiness and calm, achieving precision values of 0.91 and 0.95, respectively, with balanced recall rates. However, due to overlapping features and the inherent subtlety of these expressions, the system exhibits reduced effectiveness in identifying complex emotional states, such as frustration and stress. These findings underscore the need for additional data enrichment and advanced modeling techniques to enhance the system's adaptability in detecting nuanced emotional states across diverse educational scenarios.

1) MODEL PERFORMANCE IN EMOTION DETECTION

The system's emotional precision was evaluated by comparing the predictions generated by the model with the evaluations of emotion experts, such as psychologists and educators. During the evaluation process, data was collected on emotions detected from interactions between students and

TABLE 1. Criteria for evaluating emotions by experts (facial and vocal features).

Emotional Category	Facial Features	Vocal Characteristics
Happiness	Broad smile, bright eyes, relaxed eyebrows	High voice pitch, fast rhythm, clear and cheerful vocalization
Sadness	Downturned lips, drooping eyes, lack of brightness in eyes	Low voice pitch, slow rhythm, longer duration of vocalizations
Stress	Furrowed eyebrows, narrowed eyes, tense facial movements	Tense voice pitch, irregular rhythm, frequent pauses
Calm	Relaxed expression, neutral eyebrows, calm gaze	Soft voice pitch, constant rhythm, low and continuous volume
Frustration	Furrowed eyebrows, tight lips, uncomfortable facial movements	High voice pitch, rapidity in vocalization, more significant number of pauses
Surprise	Wide open eyes, slightly open mouth, raised eyebrows	High voice pitch, fast rhythm, abrupt changes in intonation
Interest	Raised somewhat eyebrows, focused eyes	Clear voice pitch, moderate rhythm, inflections in the voice

TABLE 2. Emotion evaluation metrics with confidence intervals (95%).

Emotion	Precision (95% CI)	Recall (95% CI)	F1-score (95% CI)
Happiness	0.91 (0.88–0.94)	0.89 (0.85–0.92)	0.90 (0.87–0.93)
Sadness	0.76 (0.73–0.80)	0.80 (0.76–0.83)	0.78 (0.75–0.81)
Stress	0.83 (0.80–0.86)	0.79 (0.76–0.82)	0.81 (0.78–0.84)
Calm	0.95 (0.92–0.97)	0.94 (0.92–0.96)	0.94 (0.92–0.96)
Frustration	0.77 (0.74–0.80)	0.74 (0.71–0.77)	0.75 (0.72–0.78)
Surprise	0.82 (0.79–0.85)	0.80 (0.77–0.83)	0.81 (0.78–0.84)
Interest	0.88 (0.85–0.91)	0.85 (0.82–0.88)	0.86 (0.83–0.89)

teachers within an educational setting. The system used voice and facial expression data to detect emotions and generated real-time predictions of students' emotions.

To evaluate the precision of these predictions, experts watched the recorded interactions. They classified the observed emotions into specific categories: happiness, sadness, stress, calm, frustration, surprise, and interest. These classifications were based on detailed observational criteria, including students' facial features and vocal responses. Each emotional category was associated with facial cues and vocal patterns that the experts used to correctly identify and classify the emotions.

Table 1 presents the evaluation criteria used by the experts to classify students' emotions during interactions in the educational setting. The experts used specific facial and vocal cues to identify each emotion, as described in the table. For example, happiness was determined by a broad smile, bright eyes, and relaxed eyebrows, accompanied by a raised voice tone, fast tempo, and clear, cheerful vocalization. In contrast, sadness was observed through a drooping facial expression and a low voice tone, with a slow tempo and longer duration of vocalizations. These facial and vocal features served as the experts' main criteria to assess.

Upon completing the manual assessment of the observed emotions, the experts compared their observations to the predictions made by the system. To make this comparison, a confusion matrix was used to measure how many predictions made by the system matched the experts' assessments (true positives), how many emotions were incorrectly classified (false positives and false negatives), and how many times the system failed to detect an emotion that was present in the interaction (false negatives). The confusion matrix provided a quantitative analysis of the system's precision. It helped identify areas where the model could improve, such as more complex and less obvious emotions.

Each emotion observed by the experts was assigned to a specific instance of the interaction and then compared to the system's corresponding prediction. The system's precision was calculated as the number of correct predictions (true positives) divided by the total predictions made for that emotion. At the same time, recall was measured as the proportion of emotional instances correctly identified by the system concerning the total emotions observed by the experts.

The results of this evaluation are shown in Table 2, which presents the precision, recall, and F1-score values for each emotion detected by the system. These results show that the system performs exceptionally well in detecting happiness and calm, with precision values of 0.91 and 0.95, respectively. These values indicate that the system can correctly identify emotions when dealing with positive or neutral emotions. The recall for these emotions is also high, with 0.89 for happiness and 0.94 for calm, suggesting that the system is accurate and capable of detecting most emotional instances associated with these emotions. The F1-score for both emotions is also notable, reflecting a balanced performance between precision and recall.

The confidence intervals (95%) for precision, recall, and F1-score were calculated using bootstrapping, a statistical resampling technique. This provides a robust measure of the variability in the metrics, offering insights into the model's stability across different test data subsets. Furthermore, sensitivity analysis was conducted to evaluate the system's performance under hardware constraints, such as low-resolution cameras (e.g., 480p) and low-fidelity microphones. These tests revealed a degradation in precision for complex emotions like frustration (a drop of 6%) and stress (a drop of 4%), emphasizing the importance of robust pre-processing techniques.

2) ANALYSIS OF CHALLENGES IN COMPLEX EMOTION DETECTION

Frustration and stress emotions underperformed compared to positive emotions. The precision for frustration was 0.77, and its recall was 0.74, indicating that the system had difficulty correctly identifying this emotion. In some cases, it also had trouble recognizing frustration in interactions. Similarly, stress had a precision of 0.83 and a recall of 0.79, suggesting that although the system is more accurate, there are still areas for improvement, as it does not always correctly detect emotions associated with stressful situations. The F1-score for frustration and stress is lower than that of positive emotions, reflecting a less balanced performance regarding precision and recall.

Including confidence intervals for these metrics provides additional insights into the variability of the system's performance. For example, the precision for frustration ranges from 0.74 to 0.80 (95% CI), and its recall ranges from 0.71 to 0.77 (95% CI). These intervals highlight the uncertainty in detecting this emotion consistently. Stress shows similar variability, with precision ranging from 0.80 to 0.86 (95% CI) and recall ranging from 0.76 to 0.82 (95% CI). These findings underscore the challenges of robustly detecting these emotions and justify the need for additional improvements.

These challenges are partly attributed to the overlap of features between these complex emotions and other categories, such as stress with frustration or sadness. Future model iterations will integrate additional training data specific to these emotions, focusing on scenarios that mimic fundamental educational interactions to capture their nuanced characteristics. Incorporating datasets like DAiSEE, designed for affective states in academic settings, can provide a broader representation of these emotions. Furthermore, creating custom datasets using real-world educational recordings will enhance the system's adaptability to context-specific emotional expressions.

Data augmentation techniques will play a critical role in addressing these challenges. For example, synthetic noise addition and pitch shifting can expand the diversity of audio data, while geometric transformations, such as rotations and zooming, will increase the variability of facial image samples. Additionally, transfer learning will leverage pre-trained models on large-scale datasets such as AffectNet, fine-tuning them with domain-specific data to improve their ability to capture subtle patterns associated with frustration and stress.

Surprise and interest emotions performed moderately, with precision values of 0.82 and 0.88, respectively, and recall of 0.80 and 0.85. Although these results are satisfactory, they indicate that the system could still benefit from adjustments to improve its detection of less evident or frequent emotions in the educational environment.

Statistical significance tests were applied to validate the observed differences in performance metrics. A Wilcoxon signed-rank test comparing F1 scores for frustration and happiness indicated significant differences ($p < 0.05$). This analysis confirms that the system performs less consistently for complex emotions like frustration than simpler ones like happiness, reinforcing the necessity of advanced modeling techniques and additional data enrichment.

Hardware limitations, such as low-resolution cameras or low-quality microphones, significantly impact the model's ability to detect subtle features required for these complex emotions. Pre-processing techniques will be employed to mitigate these issues, including noise reduction for audio and resolution enhancement for video inputs. These strategies will improve the quality of the input data, ensuring better feature extraction. Additionally, hardware recommendations will be established, suggesting minimum specifications, such as 720p camera resolution and 44.1 kHz sampling rates for microphones, to optimize performance during deployment.

The results suggest that the system effectively recognizes positive and easier-to-identify emotions, such as happiness and calm, but presents challenges when addressing more complex and subtle emotions, such as frustration and stress. These findings indicate that although the system is well-calibrated to detect certain emotions, adjustments are needed, such as increasing specific training data or using more advanced techniques to improve the model's ability to detect more varied and difficult-to-identify emotions.

Figure 4 presents the distribution graph of detected emotions, which visualizes how emotions were distributed among students during the sessions. This graph shows a high prevalence of emotions such as happiness and calm, with more than 300 detected instances of joy and around 250 of calm. In contrast, frustration and sadness were detected less frequently. This pattern, combined with the variability captured in the confidence intervals, highlights the system's strengths in identifying positive emotions while emphasizing the need for refinement in detecting complex ones like frustration and stress.

B. IMPACT OF EMOTIONS ON STUDENT PARTICIPATION

The evaluation of the impact of detected emotions on student engagement focused on analyzing how the emotional predictions generated by the system correlate with the number of student interactions during academic activities. In this regard, key metrics such as the number of responses to questions, participation in debates or forums, and the level of interaction with the content were considered. The analysis was performed by comparing the Pearson correlation values between the detected emotions and the engagement metrics to identify significant patterns that could indicate a positive or negative impact of emotions on student engagement.

3 presents the data on the correlation between the detected emotions and student engagement, along with the number of interactions and the estimated level of engagement. The results suggest that positive emotions, such as happiness (0.85) and calmness (0.90), are strongly positively correlated with engagement. These emotions are associated with a higher frequency of interactions (350 for happiness and 260 for calmness) and a high level of engagement. The system found that happier or calmer students tended to interact more, reinforcing that positive emotions foster engagement and active participation in the classroom.

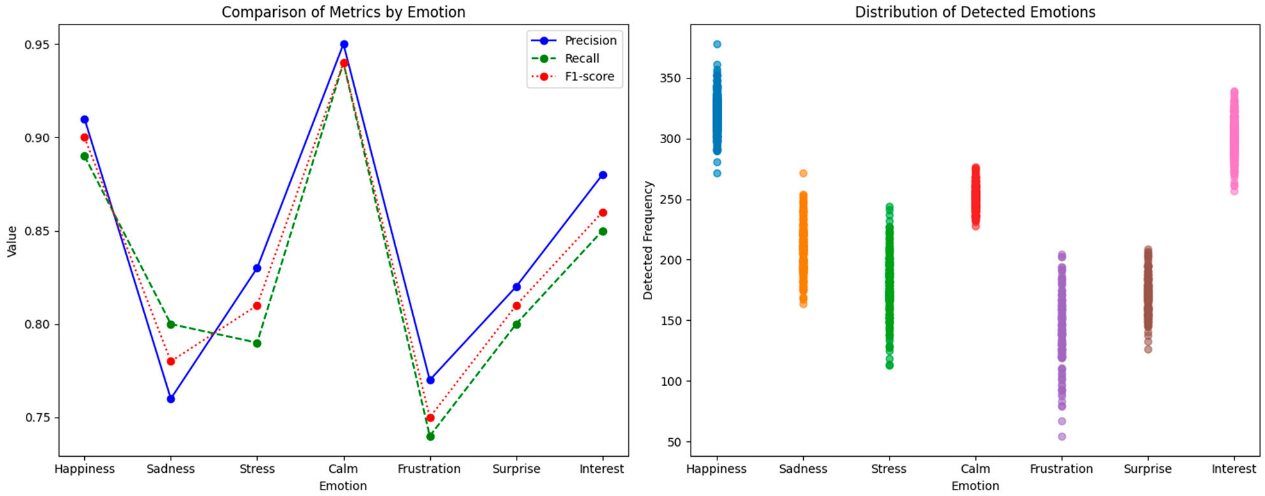

FIGURE 4. Comparison of precision, Recall, and F1-score by emotion and distribution of detected emotions. Chart A: Comparison of precision, Recall, and F1-score by emotion; Chart B: Distribution of detected emotions.

TABLE 3. Correlation between emotions and student participation.

Emotion	Correlation with Participation	Number of Interactions	Participation Level
Happiness	0.85	350	High
Sadness	-0.60	120	Low
Stress	-0.45	140	Moderate
Calm	0.90	260	High
Frustration	-0.50	150	Moderate
Surprise	0.65	180	High
Interest	0.80	300	High

In contrast, negative emotions such as stress (-0.45) and sadness (-0.60) showed negative correlations with engagement. Students who experienced these emotions showed fewer interactions (120 for sadness and 140 for stress), and the associated level of engagement was categorized as low. This pattern suggests that negative emotions may reduce students' willingness to interact, possibly related to discomfort, frustration, or emotional disconnection from the learning environment.

The analysis also included emotions such as frustration (-0.50), surprise (0.65), and interest (0.80), which showed moderate correlations with engagement. While frustration and stress were associated with lower levels of engagement, surprise and interest were positively correlated, indicating that emotions that generate curiosity or wonder can motivate engagement. In particular, the emotion of interest showed a high positive correlation (0.80), with a significant number of interactions (300), suggesting that students interested in the topic tend to participate actively.

Figure 5 shows the distribution of the detected emotions and how they relate to students' interactions. Emotions such as happiness and interest dominate the graph, showing a high frequency of participation. In contrast, feelings of stress and sadness are represented with considerably fewer interactions, confirming that negative emotions are associated with lower participation. Analyzing these results provides valuable information for optimizing the emotional detection system, allowing us to identify which emotions significantly impact participation. It can be concluded that while positive emotions such as happiness and calm clearly and strongly impact participation, negative emotions such as stress and sadness seem to inhibit interaction.

C. TEACHERS' SATISFACTION

Teacher satisfaction was assessed using Likert scale surveys, in which teachers rated various aspects of the system, such as usefulness, ease of use, effectiveness in emotional management, and contribution to classroom dynamics. These surveys provide insight into how teachers perceive the tool's impact and what aspects could be improved.

The survey design followed a structured methodology to capture teachers' perceptions comprehensively. The questionnaire consisted of 15 questions grouped into four dimensions: perceived usefulness, perceived ease of use, impact on emotional management, and contribution to classroom dynamics. Each question was rated on a 5-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). The survey was designed to align with key principles of user acceptance evaluation, focusing on capturing both quantitative and qualitative insights into the system's effectiveness and usability.

The questions were developed based on a literature review of similar emotional detection systems in educational settings, ensuring the dimensions assessed were relevant and context-specific. For instance, perceived usefulness included items such as "The system helps me better understand my students' emotions" and "The system provides actionable insights for improving classroom dynamics." Perceived ease of use addressed aspects like "The system is intuitive and requires minimal training." Experts in educational technology reviewed these questions to ensure their validity and clarity.

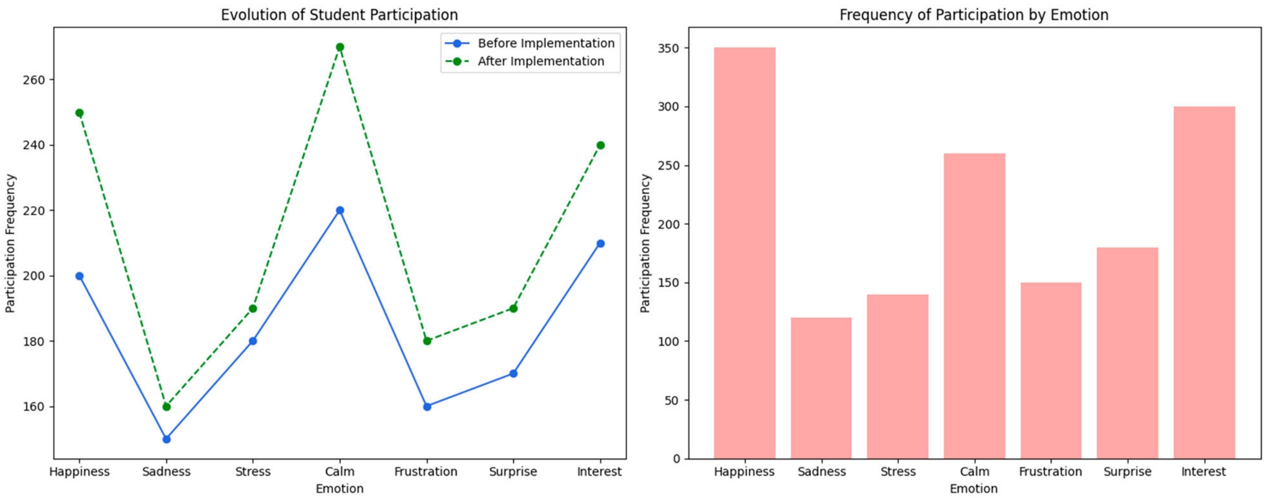

FIGURE 5. Evolution of student participation and frequency of participation by emotion. Chart A: Evolution of student participation; Chart B: Frequency of participation by emotion.

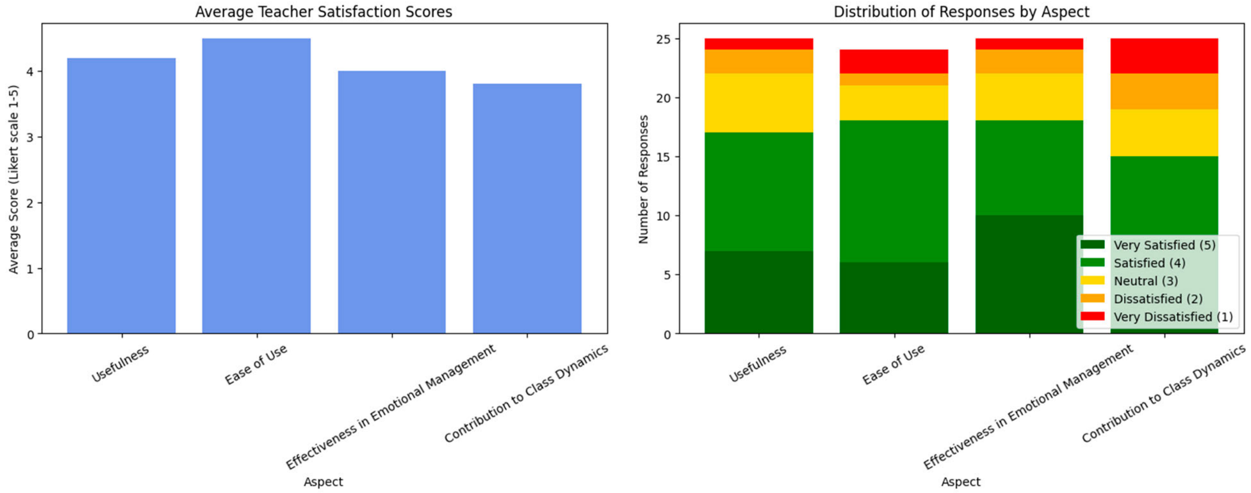

FIGURE 6. Teacher satisfaction assessment, Chart A: Average teacher satisfaction scores; Chart B: Distribution of responses by aspect.

Figure 6 presents two key representations of teacher satisfaction. The bar chart on the left shows the average of the scores for each of the aspects assessed. Teachers gave the highest scores for ease of use (4.5) and usefulness (4.2) of the system, indicating that these aspects were perceived as highly satisfactory. In contrast, contribution to classroom dynamics was given the lowest score (3.8), suggesting that although teachers consider the helpful system, its integration with classroom dynamics could be improved. This result was expected, given that the application of this type of tool in online classes still faces challenges in its effective implementation.

The stacked bar chart on the right shows the distribution of responses for each aspect, breaking down teachers' opinions into satisfaction levels. The results reflect a high proportion of positive responses for usefulness and ease of use, while less positive responses are clustered around contribution to classroom dynamics. This discrepancy highlights the need to work on the system's adaptability to the specific classroom environment and find ways to improve its impact on interactive dynamics.

Besides the Likert scale responses, open-ended questions were included to capture qualitative feedback. Teachers were asked to provide examples of how the system impacted their teaching and any challenges they faced. These responses offered more profound insights into their experiences, revealing that while most found the system helpful, some suggested improvements in the precision of emotional alerts and adaptability to diverse classroom scenarios.

While the TAM model was considered during the design phase, it was not directly implemented due to the specific context of this study, which focused on evaluating a prototype system in a controlled educational setting. Instead, the dimensions of usefulness and ease of use, core elements of

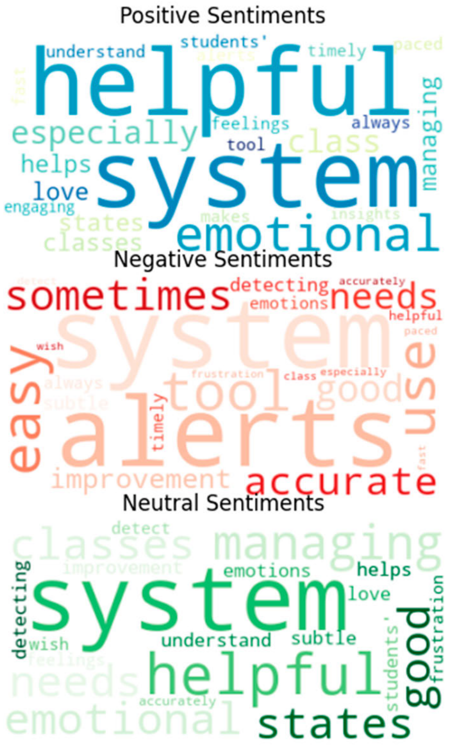

FIGURE 7. Word cloud of teacher comments.

TAM, were adapted to align with the study's goals, ensuring relevance to the educational domain without overextending the scope of the evaluation.

In addition to the quantitative results, qualitative feedback was collected from teachers about their experiences. Many highlighted that the tool helps manage students' emotions, especially identifying positive ones. However, they also suggested improving the precision of alerts for subtle or low-intensity emotions. Teachers are generally satisfied with the system's ease of use and usefulness but suggest areas for improvement in its effectiveness and adaptation to the educational environment, especially regarding its contribution to classroom dynamics.

Qualitative analysis of teachers' feedback has also offered valuable insights into their perceptions of the tool. Feedback was mostly positive, with several teachers highlighting the system's usefulness in managing students' emotions and ease of use. However, areas for improvement were also noted, especially regarding the precision of alerts and their ability to detect more complex or subtle emotions.

These comments are reflected in the word cloud presented in Figure 7, where the most common words include useful, improve, emotion, and precision. This shows that teachers are interested in improving the system and its ability to customize alerts according to students' emotional needs.

D. USABILITY AND SYSTEM EFFICIENCY

The system's usability and efficiency were evaluated to determine how teachers and students perceive the tool and how it handles response speed and data processing. This evaluation focused on critical aspects such as ease of use, response speed, and system integration within the Microsoft Teams platform, as well as measuring processing times and the generation of emotional feedback.

Table 4 shows the results of the usability surveys completed by teachers and students. Overall, the system was perceived as easy to use and valuable, with an average rating of 4.5 by teachers regarding its ease of use. Students also showed a positive rating, with an average of 4.d. The integration with Microsoft Teams received a score of 4.2 by teachers and 4.1 by students, indicating a good acceptance of the system in the virtual classroom environment.

TABLE 4. System usability survey results.

Evaluated Aspect	Teachers Average	Students Average	Number of Responses
Ease of Use	4.5	4.3	50
Speed of Response	4.0	3.8	50
Integration with Microsoft Teams	4.2	4.1	50
Efficiency in Generating Emotional Feedback	3.9	3.7	50

Despite the positive results, teachers and students indicated that emotional feedback could be improved. The teachers' score of 3.9 for the system's effectiveness in generating emotional feedback suggests that, while the system is sound, it still needs adjustments in the precision of alerts for more subtle emotions.

Teachers and students agreed that the system responded appropriately in most situations regarding response speed. However, some mentioned that more complex emotions can sometimes take longer to process. Scores in this area were 4.0 for teachers and 3.8 for students, reflecting a slight difference in perceptions of the system's speed.

Table 5 presents the average processing and emotional feedback times for each type of emotion detected. Emotional processing times were generally fast, averaging 1.1 to 1.6 seconds for emotion detection. More straightforward emotions, such as happiness and calm, had processing times

TABLE 5. System response time analysis results.

A	B	C	D
Happiness Detection	1.2	2.5	100
Stress Detection	1.5	2.7	100
Sadness Detection	1.4	2.6	100
Calm Detection	1.1	2.3	100
Frustration Detection	1.6	2.9	100
Surprise Detection	1.3	2.4	100
Interest Detection	1.2	2.5	100

Note:

- A = Scenario
- B = Average Processing Time (seconds)
- C = Average Emotional Feedback Time (seconds)
- D = Number of Processed Events

of 1.2 seconds and 1.1 seconds, respectively, demonstrating the system's efficiency in identifying clear emotions.

On the other hand, more complex emotions such as frustration and stress required more processing time, with average times of 1.6 seconds and 1.5 seconds, respectively. These longer times are understandable due to the incredible difficulty in identifying and classifying emotions that have less obvious signs. Overall, the system's response times are acceptable for real-time applications in an educational setting.

Furthermore, emotional feedback showed response times ranging from 2.3 to 2.9 seconds, depending on the emotional complexity of the response. Feedback for emotions such as happiness and calm were generated faster, while emotions such as frustration required longer times. Although the times were acceptable, these results indicate that the feedback system could benefit from further optimization to make it even more agile and adaptive.

The results obtained in terms of usability and efficiency suggest that the system is delighted by both teachers and students. The scores obtained in the survey indicate that the system is easy to use and well-integrated within Microsoft Teams, which facilitates the tool's adoption in the educational context.

However, the results also suggest areas for improvement, especially regarding the precision of emotional alerts. Teachers highlighted that, although the tool is useful, emotional alerts could be more accurate for subtle or complex emotions. In addition, feelings that are more difficult to identify, such as frustration or stress, require more processing time, reflecting the complexity of these emotions and the need to improve the system's efficiency to handle them more quickly.

Despite these areas for improvement, the system's processing and feedback times are adequate, suggesting that it is efficient in real time. This is crucial for its use in a learning environment where interaction must be dynamic and fluid.

E. STUDENT COMMENTS ON EMOTIONAL EXPERIENCE

The evaluation of student's emotional experience with the emotional detection system focused on two main areas: perception of the tool and its impact on engagement. To obtain a complete overview of the student experience, quantitative results were collected through satisfaction surveys, and qualitative results were collected through interviews and open feedback.

Table 6 presents the average scores given by students about several critical aspects of the system. The results show that students perceive the system as helpful and easy to use, with an average score of 4.4 for the usefulness of emotional feedback. Students also positively valued the integration of the system with Microsoft Teams, obtaining an average score of 4.2 for this aspect.

Despite these positive aspects, one area for improvement was highlighted: emotional monitoring. With a score of 3.9, some students expressed feelings while watched, which generated some discomfort. Although the benefits of emotional feedback were acknowledged, this perception could affect the overall experience of the system if privacy concerns are not adequately managed.

TABLE 6. Student satisfaction survey results.

Evaluated Aspect	Average Student Satisfaction	Number of Responses	Common Comment
The Usefulness of Emotional Feedback	4.4	50	Students found emotional feedback helpful in adjusting their behavior.
Perception of Emotional Monitoring	3.9	50	Some students mentioned feeling watched, although they acknowledged the usefulness.
Improvement in the Learning Experience	4.2	50	There was consensus that the system helps improve engagement and participation.
Ease of Use	4.3	50	Most students found the tool easy to use and intuitive.

In terms of effectiveness in improving the learning experience, students gave an average score of 4.2, indicating that they feel that the system helped them to participate more actively in classes. This data highlights how positive emotions such as happiness and calm are associated with increased participation, suggesting that the system positively impacts student engagement and motivation.

Table 7 presents the main comments from students regarding the use of the system. Students mentioned that one of the system's strengths is its ability to adjust their behavior and improve class participation. Comments such as "The system helped me understand my emotions and adjust my behavior" reflect how students value emotional feedback to increase classroom engagement. These comments were

TABLE 7. Students' qualitative comments on emotional experience.

Category	Common Comment	Number of Students
Strengths	The system helped me understand my emotions and adjust my behavior. I felt more motivated to participate.	30
Areas for Improvement	Sometimes, the system made me feel watched; I wanted more control over when alerts were triggered.	15
Impact on Participation	I participated more in class when I felt happy or calm, but when the system detected stress, I felt less motivated to intervene.	25
Suggestions for Improvement	It would be helpful if alerts could be more specific or personalized for each student.	20

shared among students who experienced positive emotions such as happiness or calmness.

However, areas for improvement also emerged. A significant number of students (15 in total) expressed feeling watched by the system, which can lead to discomfort. Comments such as "Sometimes the system made me feel watched" suggest that the system's sensitivity in terms of emotional monitoring should be adjusted to ensure that students do not feel intruded upon. Furthermore, students also proposed that emotional alerts could be more personalized and specific, allowing for better adaptation to individual emotional needs.

Students also noted that emotional feedback from the system directly impacted their engagement. Students generally reported that their class participation increased when they felt happy or calm, while emotions such as stress reduced their willingness to interact. This reflects how the system influences students' motivation, especially regarding their positive emotions.

As for suggestions for improvement, students proposed that alerts be made more specific for each type of emotion, allowing for a more precise response from the system. Students also suggested that greater autonomy could be offered in managing emotional monitoring, allowing students to decide when to be monitored to reduce the feeling of being watched.

F. COMPARISON WITH OTHER SOLUTIONS

The proposed emotional detection system has been compared with two popular commercial solutions in emotional analysis: Affectiva and Emotient. This subsection compares based on critical criteria such as emotional detection precision, ease of integration, scalability, and adaptability. The comparison is summarized in Table 8, and the results reflect the strengths and weaknesses of each solution based on these criteria.

TABLE 8. Comparison of the proposal with other solutions in emotional detection.

A	B	C	D
Precision in Emotional Detection	High (80%-90%)	High (85%)	Moderate (70%)
Ease of Integration	Direct integration with Microsoft Teams	Requires custom integration	Limited integration
Scalability	High (Can handle large volumes of students)	Medium (Requires tweaks for large scale)	Low (Limited to small groups)
Adaptability to the Educational Environment	High (Customizable for different educational contexts)	Medium (Requires tweaks for certain environments)	Low (Designed for specific use)

Note:

- A = Criterion
- B = Proposal
- C = Solution A (e.g., Affectiva)
- D = Solution B (e.g., Emotient)

Regarding emotional detection precision, the proposal is in the range of 80%- 90%, comparable to the precision offered by Affectiva (85%), one of the most advanced solutions on the market. However, Emotient presents a moderate precision (70%), which indicates that it does not have the same level of precision in emotional classification. These results suggest that Affectiva may be more accurate in identifying emotions, but our proposal is close to its level of precision, making it a competitive option.

An essential criterion for comparison is the ease of integration in educational environments. The proposal stands out because it integrates directly and efficiently with Microsoft Teams, a widely used distance education platform. This allows for rapid implementation without the need for technical customization. In contrast, Affectiva and Emotient require custom integrations, which can be complicated and expensive for institutions that do not have advanced technical resources. Our solution's ease of integration is a strong point that favors its adoption in diverse educational environments.

In terms of scalability, our solution demonstrates a high capacity to handle many simultaneous users or sessions, making it suitable for large virtual classrooms or educational platforms that support high demand. On the other hand, Affectiva presents medium scalability, indicating that, while it is ideal for medium-sized environments, it might require technical adjustments to function efficiently in large groups of users. Emotient shows significant limitations in terms of scalability, making it less suitable for large educational environments. This reflects that the proposal has a superior performance in terms of user volume management, a critical factor for the tool's success in large classrooms.

Adaptability is another critical factor in the comparison. The proposal is designed to be highly customizable, allowing adjustments according to the specific needs of students and the educational context. This makes it suitable for different types of classrooms and academic levels, from primary to higher education. In contrast, Affectiva and Emotient are more targeted at specific applications (e.g., market research or healthcare settings), and while they offer some flexibility, their adaptation to educational contexts is limited. This positions our solution as a more versatile option in education, where students' academic requirements vary considerably.

When comparing the proposal with Affectiva and Emotient, our solution performs competitively in terms of precision and usability and stands out in areas such as ease of integration and scalability. The ease of integration with Microsoft Teams offers a key advantage, as it allows for frictionless implementation in educational environments that already use this platform. On the other hand, Affectiva and Emotient require technical adaptations that can hinder institutions without robust technical support.

Regarding scalability, our solution has a clear advantage in handling large numbers of users without degrading performance. This is essential for use in large virtual classrooms or educational platforms supporting large student groups. Affectiva and Emotient, on the other hand, have limitations regarding the scale at which they can operate efficiently.

V. DISCUSSION

A. POTENTIAL OF EMOTIONAL DETECTION IN EDUCATION

Emotional detection in educational settings has proven to be an area with significant potential to improve student interaction, engagement, and performance, especially in virtual environments where identifying students' emotions can be difficult. Based on integrating Microsoft Teams and using advanced technologies such as TensorFlow, PyTorch, and CNN for speech and facial expression analysis, our proposal has proven effective in precision and efficiency [40]. However, it is critical to discuss the results obtained, considering previous studies, the methods used, potential limitations, and implications for future research.

Compared to commercial solutions such as Affectiva and Emotient, our tool has competitive performance in emotional precision. The works of Sin and Khin [10] and VanSteensel and Jasra [11] highlight that facial recognition and voice analysis solutions in educational settings are helpful but often lack seamless integration into popular learning platforms. While Affectiva shows high precision in emotions such as happiness and sadness, its reliance on technical customization limits its implementation in educational settings without specialized personnel. Similarly, Emotient presents precision in facial detection but also faces limitations regarding scalability in large classrooms. In this sense, our solution has been specifically designed to address these limitations, offering direct integration with Microsoft Teams and facilitating adoption in educational settings without complex technical adjustments. Furthermore, we have achieved precision comparable to commercial solutions but with the advantage of being more scalable and adaptable to varied educational contexts [41].

B. EFFECTIVENESS OF MULTIMODAL INTEGRATION

The multimodal data fusion process is another innovative aspect of our proposal. Although Hashmi and Yayilgan [13] mention that multimodal integration is critical for higher precision in emotional detection, the challenge lies in adequately combining data from different sources without losing the contextual information of each modality. Using multimodal neural networks, we have integrated voice and face signals, improving the system's ability to identify complex emotions, such as stress or frustration, which are difficult to detect with only one modality [42]. This reinforces the validity of the approach used in this work since combining visual and acoustic information is essential to obtain more accurate results in real time.

C. CHALLENGES IN COMPLEX EMOTION DETECTION

The precision in emotional detection, especially for more complex emotions, remains an area for improvement. While positive and neutral emotions like happiness and calm are easily detectable, emotions like frustration or stress present significant challenges due to their subjective and variable nature. The results in Table 1 indicate that the system had lower precision in detecting these complex emotions, with frustration achieving a precision of 0.77 and a recall of 0.74 and stress showing a precision of 0.83 and a recall of 0.79. These findings highlight the need for more sophisticated approaches to identify these emotional states precisely.

This limitation is consistent with the works of Sin and Khin [10] and VanSteensel and Jasra [11], who emphasize that emotions not explicitly associated with blatant facial expressions or marked voice timbres often require more nuanced modeling strategies. For example, frustration and stress frequently share overlapping features with other emotions, such as sadness or anger, which can lead to misclassification.

To address these challenges, future model iterations will incorporate additional training data reflecting the complexity of these emotional states. Datasets such as DAiSEE, designed explicitly to capture affective states in educational contexts, can provide more diverse and contextually relevant samples. Moreover, custom datasets derived from fundamental classroom interactions will be developed to capture nuanced variations in frustration and stress as experienced in educational settings.

Advanced techniques like transfer learning will also be leveraged to enhance the model's ability to detect these complex emotions. Pre-trained models on large-scale datasets like AffectNet will be fine-tuned using domain-specific data to improve sensitivity to subtle emotional cues. Furthermore, multi-task learning approaches will enable the model to simultaneously address related tasks, such as emotion classification and intensity estimation, which can significantly improve performance for overlapping emotional categories.

In addition to the challenges in emotion classification, hardware variability significantly impacts the system's performance. Older or low-quality webcams and microphones can introduce noise or reduce the resolution of input data, leading to degraded precision for complex emotions like frustration and stress. To mitigate these issues, preprocessing techniques such as noise reduction and resolution upscaling are implemented to improve input quality. Furthermore, edge processing capabilities enable initial data filtering and compression on the user's device, reducing reliance on high-speed internet connections and ensuring system usability even in resource-constrained environments. The system dynamically adjusts its processing pipeline based on the detected hardware capabilities, optimizing performance across diverse deployment scenarios.

Despite these challenges, the system performs well in detecting more evident emotions, such as happiness and calm, and it achieves precision and recall scores exceeding 0.90. These results indicate that the underlying framework is robust and adaptable, and the integration of advanced preprocessing and hardware-aware optimizations further strengthens its scalability in real-world applications. This foundation paves the way for further improvements through parameter tuning, additional training data, and the integration of advanced modeling techniques.

D. IMPACT ON EDUCATIONAL INTERACTION

The impact of the results on educational interaction is significant, as Zhao et al. [23] suggest that real-time analysis of students' emotions can potentially improve participation and engagement, facilitating early intervention by teachers. In this sense, the proposed system has proven helpful in providing instant emotional feedback and improving the dynamics of virtual classes. The interactivity and adaptability of the system allow teachers to adjust their pedagogical approach based on students' emotional responses, optimizing the learning experience.

E. TECHNICAL AND CONTEXTUAL LIMITATIONS

As with any AI-based system, some inherent constraints and assumptions must be discussed. First, the emotional detection model relies on pre-existing datasets such as RAVDESS for voice and AffectNet for facial expressions. Although helpful, these datasets do not fully capture the diversity of emotional expressions across different cultural and educational contexts, which could affect the system's precision when applied to students from varied backgrounds [43]. Cultural variations in facial expressions and vocal intonations introduce potential biases, as cultural norms and social interactions influence emotion perception. Future work should incorporate region-specific datasets and domain adaptation techniques to enhance the model's generalizability.

Furthermore, the system assumes that students are willing to engage in emotional monitoring, which may not always be the case. Ethical considerations related to privacy and emotional data tracking must be addressed to ensure student comfort and regulatory compliance. From a technical standpoint, the system's performance depends on hardware quality, particularly cameras and microphones. Although designed for resource efficiency, performance in environments with low-resolution audio or video recordings may be compromised, directly impacting detection accuracy. Additionally, individual emotional expression and voice tone differences could introduce model bias, limiting the system's ability to generalize across different linguistic and cultural contexts. Advanced preprocessing techniques and real-time adaptation mechanisms should be integrated into future system iterations to mitigate these issues.

VI. CONCLUSION

The study demonstrates that emotional detection in virtual educational environments, through integrating Microsoft Teams and advanced voice and facial expression analysis technologies, is an effective and scalable tool for enhancing student interaction and participation. Leveraging CNNs for facial analysis and RNNs with LSTM for voice analysis, the proposed system achieves satisfactory precision and response time, addressing several limitations of prior solutions, such as Affectiva and Emotient.

A key achievement of this work is the high precision in detecting common emotions such as happiness and calm, with values approaching 0.95. These results validate the system's effectiveness in reliably identifying positive and neutral emotional states, underscoring its potential as a practical tool to help teachers assess the overall emotional atmosphere of their classes. However, the system showed reduced precision in detecting complex emotions, such as stress and frustration, highlighting the need for further refinement to capture nuanced and negative emotional states critical in supporting students during challenging academic scenarios.

Despite these promising results, the study recognizes several limitations that must be addressed. Relying on preexisting datasets, such as RAVDESS and AffectNet, introduces potential biases in emotion detection, as these datasets may not fully reflect the cultural and linguistic diversity present in real-world educational settings. Variations in emotional expression norms may affect the system's applicability across different regions, requiring further model adaptation. Additionally, differences in student engagement and willingness to participate in emotional monitoring raise ethical considerations regarding privacy and consent.

The system's real-time emotional feedback improved classroom dynamics, enabling teachers to adapt their pedagogical strategies based on detected emotional states. However, addressing emotions such as stress and frustration requires further improvements to ensure timely and reliable emotional alerts. Fine-tuning the model's parameters could enhance these capabilities, enabling targeted interventions that foster a supportive learning environment.

Future research should focus on enhancing the system's robustness by including culturally diverse datasets and improved generalization strategies. Additionally, exploring personalized emotional feedback mechanisms would allow teachers to tailor the system's output to their students' unique emotional profiles. Transfer learning and domain adaptation techniques should also be considered to refine emotion detection across different linguistic and cultural backgrounds.

Applying emotional detection systems in professional training, remote work, or other non-educational contexts represents a valuable avenue for exploration beyond traditional educational environments. These extensions could reveal novel use cases and expand the societal impact of emotional detection technologies. Such advancements would benefit education and broader domains that rely on emotional awareness to optimize human interaction and performance.

REFERENCES

[1] W. Yang, "Extraction and analysis of factors influencing college students' mental health based on deep learning model," *Appl. Math. Nonlinear Sci.*, vol. 9, no. 1, pp. 1–14, Jan. 2024, doi: 10.2478/amns.2023.2.00773.

[2] C. Llurba, G. Fretes, and R. Palau, "Classroom emotion monitoring based on image processing," *Sustainability*, vol. 16, no. 2, p. 916, Jan. 2024, doi: 10.3390/su16020916.

[3] W. Neunzig and H. Tanqueiro, "Teacher feedback in online education for trainee translators," *Meta*, vol. 50, no. 4, pp. 1–10, Feb. 2009, doi: 10.7202/019873ar.

[4] B.-L. Jian, C.-L. Chen, M.-W. Huang, and H.-T. Yau, "Emotion-specific facial activation maps based on infrared thermal image sequences," *IEEE Access*, vol. 7, pp. 48046–48052, 2019, doi: 10.1109/ACCESS.2019.2908819.

[5] M. Roopak, S. Khan, S. Parkinson, and R. Armitage, "Comparison of deep learning classification models for facial image age estimation in digital forensic investigations," *Forensic Sci. Int., Digit. Invest.*, vol. 47, Dec. 2023, Art. no. 301637, doi: 10.1016/j.fsidi.2023.301637.

[6] T. N. Guillemette, J. L. Monn, and M. Chronister, "An evidence-based project to improve paternal postpartum depression," *J. Nurse Practitioners*, vol. 19, no. 4, Apr. 2023, Art. no. 104495, doi: 10.1016/j.nurpra.2022.11.005.

[7] C. Vogel and K. Ahmad, "Agreement and disagreement between major emotion recognition systems," *Knowl.-Based Syst.*, vol. 276, Sep. 2023, Art. no. 110759, doi: 10.1016/j.knosys.2023.110759.

[8] A. Prayogo, K. Khotimah, L. Istiqomah, and I. Maharsi, "Students' emotional engagement in online classes: A conceptual framework," *Int. J. Inf. Learn. Technol.*, vol. 41, no. 1, pp. 61–72, Jan. 2024, doi: 10.1108/ijilt-04-2023-0052.

[9] H. Shen, X. Ye, J. Zhang, and D. Huang, "Investigating the role of perceived emotional support in predicting learners' well-being and engagement mediated by motivation from a self-determination theory framework," *Learn. Motivat.*, vol. 86, May 2024, Art. no. 101968, doi: 10.1016/j.lmot.2024.101968.

[10] T. Shwe Sin and O. Khin, "Facial expressions classification on Android smartphone for a user," in *Proceedings of Sixth International Congress on Information and Communication Technology* (Lecture Notes in Networks and Systems). Singapore: Springer, Sep. 2021, doi: 10.1007/978-981-16-1781-2_21.

[11] J. VanSteensel and S. K. Jasra, "Investigating the detection of emotion concealment using the gazepoint GP3 eye-tracker," *J. Emerg. Forensic Sci. Res.*, vol. 4, no. 1, pp. 20–30, Nov. 2019.

[12] M. Lippi, M. A. Montemurro, M. Degli Esposti, and G. Cristadoro, "Natural language statistical features of LSTM-generated texts," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3326–3337, Nov. 2019, doi: 10.1109/TNNLS.2019.2890970.

[13] E. Hashmi and S. Y. Yayilgan, "Multi-class hate speech detection in the Norwegian language using FAST-RNN and multilingual fine-tuned transformers," *Complex Intell. Syst.*, vol. 10, no. 3, pp. 4535–4556, Jun. 2024, doi: 10.1007/s40747-024-01392-5.

[14] S. P. Mishra, P. Warule, and S. Deb, "Speech emotion recognition using MFCC-based entropy feature," *Signal, Image Video Process.*, vol. 18, no. 1, pp. 153–161, Feb. 2024, doi: 10.1007/s11760-023-02716-7.

[15] U. Bilotti, C. Bisogni, M. De Marsico, and S. Tramonte, "Multimodal emotion recognition via convolutional neural networks: Comparison of different strategies on two multimodal datasets," *Eng. Appl. Artif. Intell.*, vol. 130, Apr. 2024, Art. no. 107708, doi: 10.1016/j.engappai.2023.107708.

[16] M. P. Pawar, A. Rajpurohit, K. Kishor, N. Nargide, and P. More, "Backup for MS teams using graph API," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 5, pp. 7320–7326, May 2023, doi: 10.22214/ijraset.2023.53506.

[17] D. K. Nagayach and P. Verma, "Comparative analysis of big data analytics products from AWS, Azure and GCP," *IJESMS.Net*, no. 4, pp. 1450–1454, 2021.

[18] H. M. Bingen, H. I. Aamlid, B. M. Hovland, A. A. G. Nes, M. H. Larsen, K. Skedsmo, E. K. Petersen, and S. A. Steindal, "Use of active learning classrooms in health professional education: A scoping review," *Int. J. Nursing Stud. Adv.*, vol. 6, Jun. 2024, Art. no. 100167, doi: 10.1016/j.ijnsa.2023.100167.

[19] T. Zoaga Ramsa y, D. Bagdziunaite, and M. Z. Storm, "Neural predictors of ad performance, and the cannibalism of brand performance," in *Proc. NeuroPsychoEcon. Conf.*, 2015, pp. 1–18.

[20] L. Kulke, D. Feyerabend, and A. Schacht, "A comparison of the affectiva iMotions facial expression analysis software with EMG for identifying facial expressions of emotion," *Frontiers Psychol.*, vol. 11, p. 329, Feb. 2020, doi: 10.3389/fpsyg.2020.00329.

[21] A. Schmitz-Hasch, S. M. Stasch, and S. Fuchs, "Individual differences in the relationship between emotion and performance in command-and-control environments," in *Adaptive Instructional Systems. Adaptation Strategies and Methods* (Lecture Notes in Computer Science; Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Cham, Switzerland: Springer, Jul. 2021, doi: 10.1007/978-3-030-77873-6_10.

[22] E. Ortega-Ochoa, M. Arguedas, and T. Daradoumis, "Empathic pedagogical conversational agents: A systematic literature review," *Brit. J. Educ. Technol.*, vol. 55, no. 3, pp. 886–909, May 2024, doi: 10.1111/bjet.13413.

[23] M. Zhao, S. Dong, J. Hu, S. Du, C. Shi, P. Li, and Z. Shi, "Attention-guided three-stream convolutional neural network for microexpression recognition," *J. Image Graph.*, vol. 29, no. 1, pp. 111–122, 2024, doi: 10.11834/jig.230053.

[24] J. Huang, "Accelerated training and inference with the tensorflow object detection API," Google Res., Mountain View, CA, USA, 2017.

[25] N. Patel, S. Patel, and S. H. Mankad, "Impact of autoencoder based compact representation on emotion detection from audio," *J. Ambient Intell. Humanized Comput.*, vol. 13, no. 2, pp. 867–885, Feb. 2022, doi: 10.1007/s12652-021-02979-3.

[26] Y. Uranishi, "OpenCV: Open source computer vision library," *J. Inst. Image Inf. Telev. Engineers*, vol. 72, no. 9, pp. 736–739, 2018, doi: 10.3169/itej.72.736.

[27] J. J. Omena, T. Lobo, G. Tucci, E. Bitencourt, E. De Keulenaar, F. W. Kerche, J. Chao, M. Liedtke, M. Li, M. L. Paschoal, and I. Lavrov, "Quali-quanti visual methods and political bots," *J. Digit. Social Res.*, vol. 6, no. 1, pp. 50–73, Mar. 2024, doi: 10.33621/jdsr.v6i1.215.

[28] N. Chakravarty and M. Dua, "An improved feature extraction for Hindi language audio impersonation attack detection," *Multimedia Tools Appl.*, vol. 83, no. 25, pp. 66565–66590, Jan. 2024, doi: 10.1007/s11042-023-18104-9.

[29] S. Emami and V. P. Suciu, "Facial recognition using OpenCV," *J. Mobile, Embedded Distrib. Syst.*, vol. 4, no. 1, pp. 38–43, 2012.

[30] K. Chung and J.-S. Kim, "Multi-modal emotion prediction system using convergence media and active contents," *Pers. Ubiquitous Comput.*, vol. 27, no. 3, pp. 1245–1255, Jun. 2023, doi: 10.1007/s00779-021-01602-8.

[31] R. M. H. Al-Sayyed, W. A. Hijawi, A. M. Bashiti, I. AlJarah, N. Obeid, and O. Y. A. Al-Adwan, "An investigation of Microsoft Azure and Amazon Web services from users' perspectives," *Int. J. Emerg. Technol. Learn. (iJET)*, vol. 14, no. 10, p. 217, May 2019, doi: 10.3991/ijet.v14i10.9902.

[32] W. Pei, Y. Li, P. Wen, F. Yang, and X. Ji, "An automatic method using MFCC features for sleep stage classification," *Brain Informat.*, vol. 11, no. 1, p. 6, Dec. 2024, doi: 10.1186/s40708-024-00219-w.

[33] S. Mahato and S. Paul, "Analysis of region of interest (RoI) of brain for detection of depression using EEG signal," *Multimedia Tools Appl.*, vol. 83, no. 1, pp. 763–786, Jan. 2024, doi: 10.1007/s11042-023-15827-7.

[34] A. Obaigbena, O. A. Lottu, E. D. Ugwuanyi, B. S. Jacks, E. O. Sodiya, O. D. Daraojimba, and O. A. Lottu, "AI and human–robot interaction: A review of recent advances and challenges," *GSC Adv. Res. Rev.*, vol. 18, no. 2, pp. 321–330, Feb. 2024, doi: 10.30574/gscarr.2024.18.2.0070.

[35] P. S. Tomar, K. Mathur, and U. Suman, "Fusing facial and speech cues for enhanced multimodal emotion recognition," *Int. J. Inf. Technol.*, vol. 16, no. 3, pp. 1397–1405, Mar. 2024, doi: 10.1007/s41870-023-01697-7.

[36] M. K. Singh, "Multimedia application for forensic automatic speaker recognition from disguised voices using MFCC feature extraction and classification techniques," *Multimedia Tools Appl.*, vol. 83, no. 32, pp. 77327–77345, Feb. 2024, doi: 10.1007/s11042-024-18602-4.

[37] G. N. Ambika and Y. Suresh, "Mathematics for 2D face recognition from real time image data set using deep learning techniques," *Bull. Electr. Eng. Informat.*, vol. 13, no. 2, pp. 1228–1237, Apr. 2024, doi: 10.11591/eei.v13i2.5424.

[38] M. L. McHugh, "Multiple comparison analysis testing in ANOVA," *Biochemia Medica*, vol. 21, no. 3, pp. 203–209, Oct. 2011, doi: 10.11613/bm.2011.029.

[39] B. Li, M. Xu, Y. Zhou, H. Liu, and R. Zhang, "Optimization of security identification in power grid data through advanced encryption standard algorithm," *J. Cyber Secur. Mobility*, vol. 13, no. 2, pp. 239–264, Feb. 2024, doi: 10.13052/jcsm2245-1439.1323.

[40] A. W. Ou, C. Stöhr, and H. Malmström, "Academic communication with AI-powered language tools in higher education: From a post-humanist perspective," *System*, vol. 121, Apr. 2024, Art. no. 103225, doi: 10.1016/j.system.2024.103225.

[41] M. Moreno, R. Schnabel, G. Lancia, and E. Woodruff, "Between text and platforms: A case study on the real-time emotions & psychophysiological indicators of video gaming and academic engagement," *Educ. Inf. Technol.*, vol. 25, no. 3, pp. 2073–2099, May 2020, doi: 10.1007/s10639-019-10031-3.

[42] T. Gerosa, G. Argentin, and A. Spada, "What are teacher relational skills? A defining study using a bottom-up modified delphi method," *Qual. Quantity*, vol. 58, no. 1, pp. 581–602, Feb. 2024, doi: 10.1007/s11135-023-01638-3.

[43] W. Simms and M. Shanahan, "Qualitatively recognizing the dimensions of Student environmental identity development within the classroom context," *J. Res. Sci. Teaching*, vol. 61, no. 1, pp. 3–37, Jan. 2024, doi: 10.1002/tea.21863.

WILLIAM VILLEGAS-CH (Member, IEEE) received the master's degree in communications networks and the Ph.D. degree in computer science from the University of Alicante. He is currently a Systems Engineer specializing in robotics in artificial intelligence. He is a Professor of information technology with the Universidad de Las Américas, Quito, Ecuador. He has participated in various conferences as a speaker on topics, such as ICT in education and how they improve educational quality and student learning. His main articles focus on the design of ICT systems, models, and prototypes applied to different academic environments, especially with the use of big data and artificial intelligence as a basis for creating intelligent educational environments. His main research interests include web applications, data mining, and e-learning.

ROMMEL GUTIERREZ received the master's degree in cybersecurity. He is currently a Research Technician with UDLA, Quito, Ecuador, where he applies his knowledge in software development, data science, and cybersecurity, as an IT Engineer. His focus on AI, data science, cybersecurity, and software development are mainly geared towards education and research. He's passionate about utilizing these technological tools to fortify digital systems and create innovative solutions, emphasizing their applicability in educational settings and research environments.

ARACELY MERA-NAVARRETE received the master's degree in business administration from UIDE. She is currently a Computer Engineer in Quito, Ecuador. She is an Expert in E-learning platforms FATLA.ORG, her skills and abilities are in computer science and its associated technologies, such as hardware, software, communications, e-learning platforms, construction of computer systems, and management in LMS applications (Moodle-CANVAS).