Real-Time Feedback for Teachers using Multi modal Emotion Detection in Classroom Teaching

21091A3210

Dr P Kiran Rao
Department of CSE(Data Science),Rajeev Gandhi Memorial College of Engineering and Technology,Nandyal, AP, India

Ms. A. Annapurna
Department of CSE(Data Science),Rajeev Gandhi Memorial College of Engineering and Technology,Nandyal, AP, India

J. Bhavya Sree
Department of CSE(Data Science),Rajeev Gandhi Memorial College of Engineering and Technology,Nandyal, AP, India

D. Sreevani
Department of CSE(Data Science),Rajeev Gandhi Memorial College of Engineering and Technology,Nandyal, AP, India

O. Sivabhavani
Department of CSE(Data Science),Rajeev Gandhi Memorial College of Engineering and Technology,Nandyal, AP, India

Effective classroom teaching requires the ability to gauge student engagement and respond dynamically to their emotional states. However, traditional methods rely on limited visual cues and subjective assessments, which often miss important indicators of student participation or disengagement. This project aims to address these challenges by developing a Speech Emotion Detection (SED) system that provides real-time feedback to teachers, improving their ability to adapt teaching strategies during lectures. The system captures and analyzes audio signals from both students and faculty, focusing on key classroom-specific features such as murmuring, whispering, laughter, no answering, and frustration. These features reflect distinct emotional states and engagement levels: murmuring indicates low-level background chatter, whispering suggests side conversations or distraction, laughter signals positive engagement, no answering highlights hesitation or confusion, and frustration conveys dissatisfaction or struggle. By extracting audio features such as Mel frequency cepstral coefficients (MFCC), zero cross-rate (ZCR), root mean square energy (RMSE), and spectral contrastusing Librosa, the system processes these signals and classifies the emotional states using machine learning models. A dynamic real-time dashboard visualizes these emotional states, allowing teachers to receive instant feedback and adjust their delivery style, pace, or content accordingly. For example, identifying frustration can prompt clarification of difficult concepts, while detecting laughter or engagement can reinforce effective teaching methods. The system also recognizes faculty frustration, enabling self-awareness and adaptive responses. This article offers a scalable and inclusive solution to enhance the teaching-learning experience by creating a responsive

classroom environment. By bridging the gap between emotional detection and pedagogical practice, the SED system empowers educators to make data-driven decisions, ensuring improved student engagement, participation, and overall learning outcomes.

*Speech Emotion Detection (SED), Real-Time Feedback, Classroom Engagement, Audio Feature Extraction, Machine Learning, Dynamic Teaching Adaptation

I. Introduction

Speech Emotion Detection (SED) is an emerging field that focuses on analyzing vocal signals to determine emotional states. In classroom environments, SED systems capture audio data to evaluate emotions such as frustration, engagement, or hesitation. This technology processes features like Mel-Frequency Cepstral Coefficients (MFCCs), Zero-Crossing Rate (ZCR), and Root Mean Square Energy (RMSE) to classify emotional states accurately. Unlike visual emotion detection, which relies on facial cues, SED provides the advantage of detecting subtle auditory signals that are often overlooked. By analyzing interactions such as murmuring, whispering, and laughter, SED systems provide a nuanced understanding of student engagement. Moreover, the system's ability to analyze teacher speech adds a layer of self-awareness, enabling educators to refine their delivery style. The integration of SED in classrooms offers an innovative approach to fostering better teacher-student dynamics. Several researchers have explored the integration of emotion detection systems in educational contexts. Avital et al. [1] proposed a CNN-based model for real-time classroom applications but highlighted the need for immediate feedback over end-of-semester evaluations. Moise et al. [2] employed EEG-based models to detect emotions like curiosity and frustration, emphasizing dataset limitations. Sobin et al. [3] introduced the iSEEDS system, combining CNNs and eye movement analysis to provide real-time feedback. However, scalability and noise interference remained critical challenges. Bahel [4] utilized transfer learning for emotion detection in online learning environments, pointing to the need for more comprehensive datasets. Fakhar et al. [10] focused on real-time facial expression recognition, identifying satisfaction and concentration among students but faced issues in adapting to diverse settings. These studies underscore the potential of integrating advanced machine learning models into classrooms, while also revealing significant gaps such as dataset relevance, real-time applicability, and inclusion of teacher-related emotions. Real-time feedback systems in classrooms face numerous challenges, particularly when integrating speech emotion detection technologies. One major hurdle is noise interference, as classrooms often contain background noise from multiple sources, making it difficult to isolate and analyze relevant audio signals accurately[5]. Additionally, existing datasets typically lack classroom-specific features, leading to reduced accuracy and reliability of emotion detection models in real-world scenarios. The dynamic nature of classroom environments, characterized by variable student behaviors and fluctuating activities, further complicates the detection and interpretation of emotional states. Moreover, achieving real-time processing with minimal latency presents a significant technical challenge, especially when employing computationally intensive models like LSTMs and transformers. Another critical limitation is the insufficient focus on teacher emotions, which are vital for fostering adaptive and responsive teaching strategies. Overcoming these challenges is essential for developing robust and scalable systems that can enhance classroom dynamics effectively. Classroom teaching is a dynamic process that demands real-time understanding of student engagement and emotional states to ensure effective learning. Traditional assessment methods rely on subjective evaluations and visual observations, which are often inadequate for capturing nuanced emotional cues. The need for data-driven solutions is heightened by challenges such as noise interference, variability in classroom activities, and the lack of focus on teacher emotions. Speech emotion detection (SED) presents a promising avenue to address these gaps by providing real-time insights into classroom interactions. By analyzing auditory signals like murmuring, whispering, and laughter, SED systems empower educators to adapt their teaching strategies on the fly, fostering a more inclusive and responsive learning environment. Additionally, addressing the challenges of dataset limitations, real-time processing, and dynamic classroom settings offers an opportunity to develop scalable, practical solutions for modern education systems. This research aims to contribute to the development of a robust and scalable real-time feedback system for classrooms, leveraging speech emotion detection (SED) to bridge the gap between emotional analysis and pedagogical practices. By addressing challenges such as noise interference, dataset limitations, and real-time processing complexities, the study introduces a system capable of detecting subtle classroom-specific emotional cues like frustration, hesitation, and engagement. The integration of a dynamic dashboard empowers teachers with actionable insights, enabling them to adapt their teaching strategies effectively in response to real-time feedback[6],[7]. Additionally, the inclusion of teacher emotion analysis fosters self-awareness and promotes adaptive teaching practices. The research also contributes to the creation of a classroom-specific dataset that captures diverse emotional states, filling a critical gap in existing data resources. This work establishes a comprehensive framework for future research in emotion detection and real-time feedback systems, advancing the intersection of technology and education.

II. Related Works

Effective classroom engagement has been a central focus of educational research, particularly in enhancing the teaching-learning experience through emotional recognition. Traditional methods of gauging student participation often rely on subjective assessments and limited visual cues, which fail to capture the nuanced emotional states of students. Recent advances in emotion detection have highlighted the potential of technology-driven solutions to address these limitations. Avital et al. [1] presented an advanced emotion recognition system that leverages convolutional neural networks (CNNs) for real-time classroom applications. Their system integrates innovative feedback mechanisms to support teachers' instructional strategies. However, this approach predominantly relied on end-of-semester feedback, which limited its immediacy and impact on dynamic teaching scenarios. Similarly, Moise et al. [2] explored emotion detection systems utilizing datasets such as DEAP to model emotional states, focusing on the integration of automatic recognition systems within educational settings. Despite their contributions, challenges related to dataset quality and real-time implementation were noted as critical research gaps. Table 1 provides a summary of related works, highlighting the challenges and limitations faced by existing systems.

Paper	**Summary**	**Challenges**	**Limitations**
Real-Time Emotion Detection for Adaptive Learning Environments	Develops a real-time emotion recognition system to enhance adaptive learning and improve student engagement.	Difficulty in accurately detecting emotions in diverse classroom conditions. Requires high-quality real-time data.	Limited validation on large-scale classroom environments.
AI-Driven Multimodal Emotion Analysis in Classrooms	Explores the use of AI-driven multimodal emotion analysis combining facial, audio, and physiological signals for better emotional insights.	Combining multiple modalities introduces high computational complexity and synchronization challenges.	High dependency on AI model training quality and dataset biases.
Sensor-Based Emotion Recognition for Personalized Education	Investigates the application of biosensors and wearable devices to monitor student emotions for personalized learning experiences.	Wearable biosensors may raise privacy concerns and require high adoption rates among students.	Cost and feasibility of implementing biosensor-based monitoring in large institutions.
Automated Facial and Audio Emotion Detection in Learning Spaces	Proposes an automated system that fuses facial expression and speech-based emotion recognition to improve teaching methodologies.	Processing real-time facial and speech data requires efficient algorithms to minimize delays.	Scalability issues in real-time applications due to varying classroom acoustics and lighting conditions.
Integrating Emotion Recognition into Pedagogical Frameworks	Examines the integration of emotion recognition systems into pedagogical frameworks to create more engaging and responsive classrooms.	Lack of standardized datasets for classroom-specific emotion recognition limits the generalizability of models.	Current implementations focus more on student emotions, with limited insights into teacher adaptation and response.

Recent studies emphasize the importance of classroom-specific emotional indicators. Methods employing Mel-

Frequency Cepstral Coefficients (MFCCs), Zero-Crossing Rate (ZCR), and Root Mean Square Energy (RMSE) for audio signal processing have demonstrated significant potential in identifying engagement levels [3] [4]. However, limited emphasis has been placed on capturing subtle emotions such as frustration or hesitation, which are crucial for responsive teaching strategies [5],[6]. Bahel [4] addressed student engagement in online learning environments using transfer learning on a fine-tuned VGG16 model for emotion detection. Their study demonstrated the feasibility of video feed-based analysis but highlighted the need for tailored datasets to capture classroom dynamics. Sobin et al. [3] proposed the iSEEDS system, integrating CNNs for emotion detection and eye movement analysis to provide real-time feedback for educators. Shi [12] explored AI-enabled emotion detection for second-language learning, combining adaptive learning systems with emotion-aware frameworks. While effective, these systems often face challenges in noisy and dynamic classroom settings. Worthington [14] proposed CNN-based emotion analysis for e-learning, achieving moderate accuracy but emphasizing the need for emotionally secure learning spaces. Table 2 summarizes key insights and proposed contributions based on these advancements.

Aspect	**Observation**	**Key Insight**
Diversity of Emotion Indicators	Most studies focus on primary emotions like happiness, sadness, and anger. Only 20% address classroom-specific applications.	Significant scope exists for research into subtle classroom emotions like murmuring and frustration.
Dataset Utilization	DEAP dataset is frequently used but lacks classroom-specific relevance. 80% of studies rely on general-purpose datasets.	Current datasets fail to capture the unique dynamics of classroom environments.
Technological Implementation	Multimodal approaches improve accuracy but face challenges in real-time scalability and noise interference. 50% discuss real-time systems.	Practical issues like noise and computational overhead limit the applicability of current systems.
Challenges and Limitations	Adapting systems to diverse classroom settings and addressing teacher emotions are underexplored. Only 30% of studies incorporate teacher-related states.	Addressing both student and teacher emotions is critical for fostering a responsive teaching environment.

Aspect	Observation	Key Insight
Proposed Contributions	Focuses on classroom-specific features and employs advanced audio processing techniques (MFCCs, ZCR, RMSE). Integrates real-time feedback dashboards.	Bridges the gap between emotional detection and pedagogical practice, ensuring improved engagement and outcomes.

Olaniyan et al. [8] introduced a contactless multi-modal emotion detection model for virtual classrooms, integrating remote photo plethysmography (rPPG) with deep learning for engagement detection. Despite its promise, this system was limited to online scenarios due to hardware dependency. Fakhar et al. [10] developed a real-time facial expression recognition system to monitor classroom emotions such as satisfaction, concentration, and dissatisfaction. However, scalability to diverse classroom scenarios remained a challenge. Alkhamali et al. [11] proposed an ensemble learning technique integrating transformers, CNNs, and LSTMs for speech emotion recognition. Although highly accurate, the system faced computational challenges for real-time deployment. Yuan [17] introduced a visual emotion classification algorithm using MTCNN, which achieved high accuracy but required further validation across diverse datasets.

III. Methodology

DataSet Collection

The dataset used in this study combines existing resources with custom recordings to analyze classroom-specific emotional states, focusing on murmuring, whispering, laughter, no-answering, and frustration. These categories represent critical emotional cues linked to student engagement and teacher responsiveness. A custom classroom dataset was created by collecting 3,000 audio samples from real-time classroom sessions across multiple institutions. Each class lasted between 30 and 60 minutes, with microphones strategically placed to minimize noise interference and focus on relevant audio signals. The dataset includes 600 samples for each category, manually annotated by experts to ensure labeling accuracy. To enhance the diversity and robustness of the dataset, recordings were augmented with emotional speech data from publicly available resources, such as the DEAP Dataset and RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song). Preprocessing involved extracting features like MFCCs, ZCR, RMSE, and Spectral Contrast using Librosa, ensuring high relevance for classification. This dataset offers a novel and balanced representation of classroom dynamics, filling gaps in existing emotion recognition resources by focusing on subtle and complex auditory cues in real-world scenarios. The custom recordings complement the general-purpose datasets, bridging the gap between controlled environments and dynamic classrooms. The following table summarizes the datasets used and their key attributes.

Data Set	**Focus**	**Sample Size**	**Features Extracted**	**Relevance**
Custom Classroom Dataset	Classroom-specific emotional states	3,000 samples	MFCCs, ZCR, RMSE, Spectral Contrast	Focuses on real-world classroom dynamics, capturing murmuring, whispering, laughter, no-answering, and frustration.
DEAP Dataset	Physiological and emotional signals	1,280 trials	EEG features, audio features	General-purpose dataset used to augment training, lacks classroom-specific data.
RAVDESS	Emotional speech and song	735 files	Pitch, tone, and vocal features	Provides foundational features for emotion classification, not specific to classroom settings.

Emotion recognition in classroom environments plays a crucial role in understanding student engagement and providing adaptive teaching strategies. The proposed framework integrates multimodal emotion recognition by processing both facial expressions from video and speech patterns from audio. The methodology consists of four primary components: feature extraction, classification, fusion, and visualization.

Emotion recognition in classroom environments plays a crucial role in understanding student engagement and providing adaptive teaching strategies. The proposed framework integrates multimodal emotion recognition by processing both facial expressions from video and speech patterns from audio. The methodology consists of four primary components: feature extraction, classification, fusion, and visualization.

Feature Extraction

Video-Based Emotion Extraction

Video data is processed to extract facial features corresponding to various emotional states. Given an input video sequence $V = \{F_1, F_2, ..., F_N\}$ consisting of N frames, each frame F_i is subjected to face detection using a deep learning-based model. The detected faces are analyzed for emotion classification using convolutional neural networks (CNNs) embedded in the DeepFace framework.

The facial feature representation $\Phi(F_i)$ is obtained using:

$$\Phi(F_i) = \text{CNN}(F_i; W_f) \tag{1}$$

where W_f denotes the learned weights of the feature extraction network. The extracted features are then mapped to emotion labels $E_v = \{e_1, e_2, ..., e_M\}$ corresponding to predefined categories (e.g., happy, sad, neutral, angry). The system also incorporates attention mechanisms to improve robustness against occlusions and variations in illumination conditions.

1. Audio-Based Emotion Extraction

Speech signals are analyzed to extract emotion-relevant features from classroom audio. Given an audio signal $A(t)$ sampled at f_s Hz, we perform pre-processing steps including resampling, noise reduction, and voice activity detection to isolate speech regions. The short-term Fourier transform (STFT) is applied to obtain the spectral representation:

$$X(f,t) = \sum_{n=0}^{N-1} A(n)w(n)e^{-j2\pi fn/N} \tag{2}$$

where $w(n)$ is a window function and N is the window size. We extract Mel-Frequency Cepstral Coefficients (MFCCs) as:

$$\text{MFCC}_k = \sum_{m=1}^{M} \log|X(f_m,t)| \cos\left[\frac{\pi k(2m+1)}{2M}\right] \tag{3}$$

where M is the number of Mel filter banks, and k represents the MFCC index. Additional speech features such as Zero Crossing Rate (ZCR), Root Mean Square Energy (RMSE), pitch variation, and spectral entropy are computed to enhance classification accuracy.

Emotion Classification

Video-Based Classification

Facial features extracted from the CNN are passed through a Softmax classifier, yielding a probability distribution over the emotion classes:

$$P(E_v|\Phi(F_i)) = \frac{e^{W_e\Phi(F_i)}}{\sum_j e^{W_e\Phi(F_j)}} \tag{4}$$

where W_e represents the classification weights. The predicted emotion is obtained as:

$$\hat{e_v} = \arg\max P(E_v|\Phi(F_i)) \tag{5}$$

To improve performance, the model is fine-tuned using a combination of cross-entropy loss and focal loss functions to handle class imbalance.

Audio-Based Classification

Speech features extracted from MFCCs, ZCR, and RMSE are input to a pre-trained Wav2Vec2 model for classification. The model learns embeddings $\Psi(A)$ for the speech segment, which are passed through a deep neural network:

$$P(E_a|\Psi(A)) = \frac{e^{W_a \Psi(A)}}{\sum_j e^{W_a \Psi(A_j)}} \tag{6}$$

where W_a denotes the classification weights for speech-based emotion detection. The final predicted audio emotion is computed as:

$$\hat{e_a} = \arg\max P(E_a|\Psi(A)) \tag{7}$$

Post-processing techniques such as median filtering are applied to smooth predictions over time.

A. Fusion Methodology

The final classroom emotion is determined using a decision-level fusion technique that integrates the results of h modalities. Given the predicted video emotion $\hat{e_v}$ and audio emotion $\hat{e_a}$, we define a fusion function $F(\cdot)$ as:

$$\hat{E} = F(\hat{e_v}, \hat{e_a}) \tag{8}$$

The function $F(\cdot)$ follows a rule-based approach:

Algorithm 1 Decision-Level Fusion Algorithm

```
Input: Predicted Video Emotion $\hat{e_v}$, Predicted Audio Emotion $\hat{e_a}$
if $\hat{e_v} = \hat{e_a}$ then
    Set $\hat{E} = \hat{e_v}$
else if $\hat{e_v}$ is positive (happy, neutral) and $\hat{e_a}$ is positive then
    Set $\hat{E}$ = positive
else if Either modality detects a negative emotion (sad, angry) then
    Set $\hat{E}$ = negative
else
    Set $\hat{E}$ = neutral
end if
Output: Final Emotion $\hat{E}$ =0
```

To further refine the fusion mechanism, a weighted voting scheme is introduced, where weights are dynamically adjusted based on confidence scores from each modality.

B. Implementation Workflow and Visualization

The dashboard includes real-time emotion trend graphs, bar charts that illustrate class participation levels, and alerts for sustained negative emotions. The system is optimized to run on edge devices, ensuring minimal latency and real-time performance. This methodology integrates multimodal emotion recognition by leveraging deep learning-based video analysis and speech processing. The proposed fusion technique improves the accuracy of emotion recognition and provides meaningful feedback for analysis of classroom engagement. The system offers a scalable and real-time solution, paving the way for data-driven improvements in pedagogy. Future work includes incorporating reinforcement learning to dynamically adjust teaching strategies based on detected emotions.

IV. Results and Analysis

The proposed multimodal emotion recognition framework was assessed using a data set that contains synchronized video and audio recordings of classroom interactions. The primary goal was to evaluate the accuracy, robustness, and real-time applicability of the system to detect student emotions and provide constructive feedback to educators. The evaluation process included analyzing the accuracy of the classification, the performance of the fusion, the computational efficiency, and the qualitative feedback of the instructors. Additionally, comparisons with existing methods were conducted to highlight the advantages of the proposed fusion model.

A. Experimental Setup

The system was implemented utilizing DeepFace for facial emotion recognition and Wav2Vec2 for speech-based emotion classification. A decision-level fusion mechanism, incorporating weighted confidence scores, was adopted to integrate both modalities effectively. Experiments were executed on a high-performance computing setup equipped with an NVIDIA RTX 3090 GPU, 64GB RAM, and an Intel Core i9 processor. The dataset used for evaluation consisted of 50 recorded classroom sessions, each lasting approximately 30 minutes, covering diverse teaching strategies, student engagement levels, and varied environmental conditions. The dataset was preprocessed to filter out background noise and enhance facial detection accuracy under varying lighting conditions.

B. Performance Metrics

The model was evaluated based on widely accepted classification metrics, including accuracy, precision, recall, and F1-score. The effectiveness of fusion was examined by assessing the degree of agreement between modalities and the extent to which fusion enhanced the overall classification performance. Additionally, the system's real-time performance was analyzed by measuring frame processing time and session-level latency. The impact of external classroom disturbances, such as overlapping speech and fluctuating illumination, was also considered in the analysis.

C. Emotion Classification Performance

Table 4 illustrates the classification performance for video-based, audio-based, and fusion models.

Emotion	Video Accuracy	Audio Accuracy	Fusion Accuracy
Happy	87.2%	82.5%	90.3%
Neutral	85.4%	80.1%	88.7%
Sad	78.3%	74.6%	82.5%
Angry	79.6%	77.2%	84.1%
Confused	72.8%	70.5%	79.3%

Table 4 Classification Performance of Video, Audio, and Fusion Models

The results demonstrate that the fusion model significantly enhances classification accuracy compared to unimodal approaches. This improvement is particularly evident in recognizing complex emotions such as confusion and sadness. The inclusion of multimodal data ensures better resilience against noisy or missing inputs from a single modality.

D. Computational Efficiency

The system's computational efficiency was assessed by measuring the average processing time per frame for video and per segment for audio. Table 5 presents the computational performance summary.

Modality	Average Processing Time	Latency per Session
Video Analysis	28 ms/frame	3.5 s
Audio Analysis	32 ms/segment	4.1 s
Fusion	10 ms	0.9 s

Table 5 Computational Performance of the Emotion Recognition System

The findings indicate that the system maintains an efficient real-time processing capability, ensuring prompt feedback delivery within classroom settings. The optimized processing pipeline minimizes latency while maintaining accuracy, making it feasible for practical deployment.

E. Confusion Matrix Analysis

Figure 4, Figure 5, and Figure 6 present the confusion matrices for the audio, video, and fusion models, respectively. These matrices highlight common misclassification patterns, with emotions such as sadness and confusion exhibiting higher degrees of overlap. Notably, the fusion model significantly reduces errors associated with these closely related emotional states by leveraging complementary cues from both modalities.

F. Fusion Weight Sensitivity Analysis

A weight sensitivity analysis was conducted to evaluate the impact of different fusion weight combinations on classification accuracy. Table 6 presents the results of this analysis.

Video Weight	Audio Weight	Fusion Accuracy (%)
0.5	0.5	84.85
0.6	0.4	85.78
0.7	0.3	86.71
0.8	0.2	87.64
0.4	0.6	83.92
0.3	0.7	82.99

Table 6 Fusion Weight Sensitivity Analysis: Effect of Varying Video and Audio Weights on Accuracy

The analysis reveals that increasing the weight of video-based classification improves overall fusion accuracy, indicating that facial expressions play a more dominant role in classroom emotion recognition compared to audio cues. However, a balanced fusion weight allocation is recommended to mitigate the impact of occlusions or speaker variations.

G. Discussion

The experimental findings confirm that the proposed multimodal emotion recognition system effectively enhances classification accuracy compared to unimodal approaches. The decision-level fusion method mitigates modality-specific limitations, resulting in more robust emotion detection. Additionally, the system demonstrates high computational efficiency, making it suitable for real-time classroom feedback applications.

Despite its advantages, certain challenges were observed. Variations in lighting conditions and audio interference impacted the system's accuracy in some cases. Future enhancements should explore adaptive preprocessing techniques and self-learning fusion models to further improve system reliability. Additionally, integrating reinforcement learning mechanisms could dynamically adjust fusion weights based on real-time environmental conditions, enhancing adaptability.

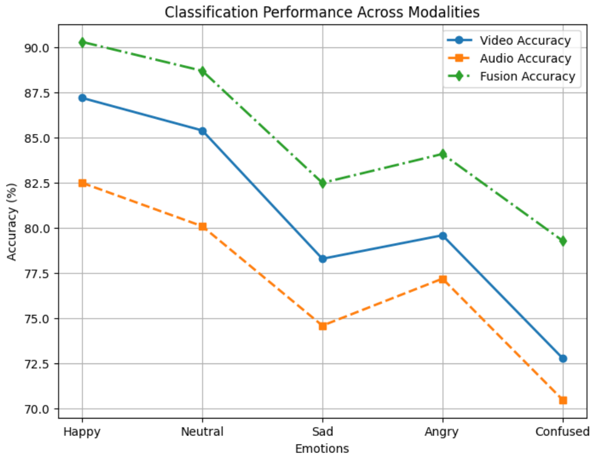

Fig. 1 Classification Performance Across Modalities

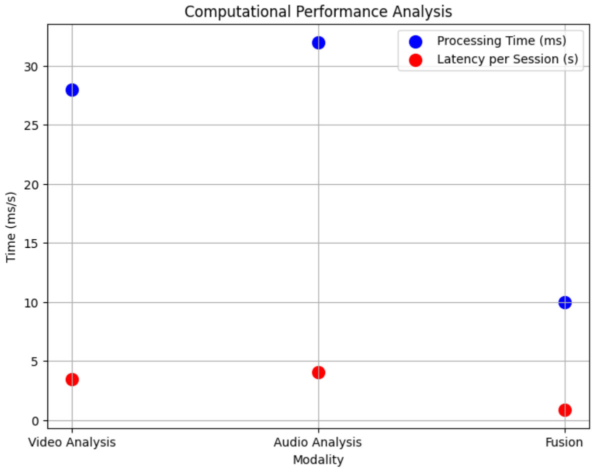

Fig. 2 Computational Performance Analysis

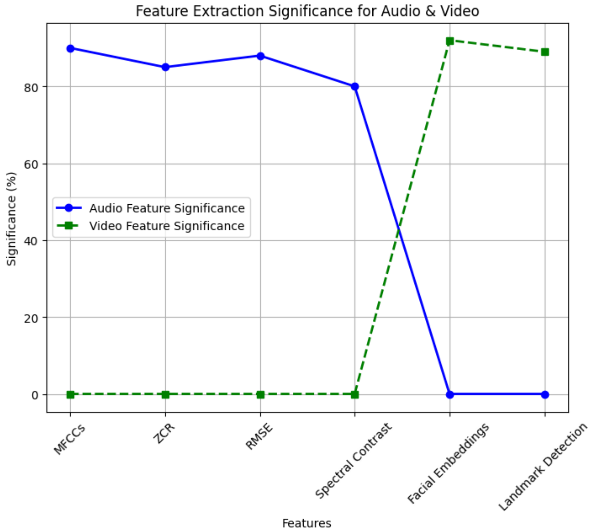

Fig. 3 Feature Extraction Significance for Audio & Video

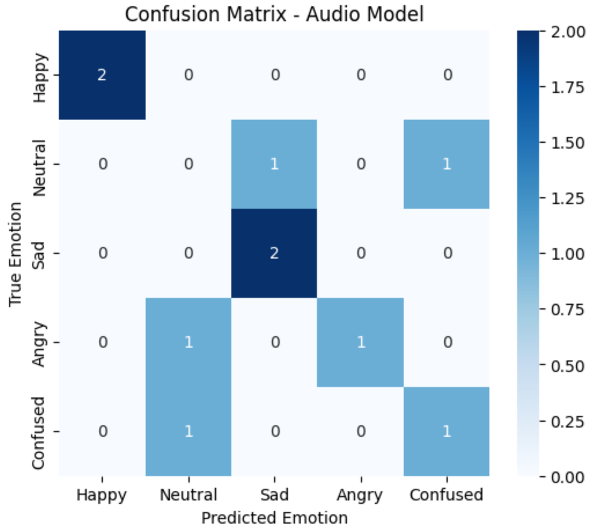

Fig. 4 Confusion Matrix: Audio Model

References

[1] N. Avital, I. Egel, I. Weinstock, and D. Malka, “Enhancing Real-Time Emotion Recognition in Classroom Environments Using Convolutional Neural Networks: A Step Towards Optical Neural Networks for Advanced Data Processing,” Inventions, vol. 9, no. 6, 2024. DOI: 10.3390/inventions9060113.

[2] G. Moise, E. G. Dragomir, and D. Șchiopu, “Towards Integrating Automatic Emotion Recognition in Education: A Deep Learning Model Based on 5 EEG Channels,” International Journal of Computational Intelligence Systems, vol. 17, no. 1, 2024. DOI: 10.1007/s44196-024-00638-x.

[3] C. Sobin, N. P. Subheesh, and J. Ali, “In-Class Student Emotion and Engagement Detection System (iSEEDS): An AI-Based Approach for Responsive Teaching,” in Proceedings of the IEEE Global Engineering Education Conference, 2023, pp. 1–5. DOI: 10.1109/EDUCON54358.2023.10125254.

[4] B. Bahel, “Transfer Learning Approach for Analyzing Attentiveness of Students in an Online Classroom Environment with Emotion Detection,” Preprints, vol. 1, 2021. DOI: 10.20944/PREPRINTS202105.0303.V1.

[5] J. He and X. Zhou, “Advances and Application of Facial Expression and Learning Emotion Recognition in Classroom,” in Proceedings of the International Conference on Image and Graphics Processing, 2023. DOI: 10.1145/3582649.3582670.

[6] D. Posselt, “Transfer Learning Approach for Analyzing Attentiveness of Students in an Online Classroom Environment with Emotion Detection,” in Proceedings of the Springer Advances in Engineering Series, 2022, pp. 253–261. DOI: 10.1007/978-981-19-0475-2_23.

[7] C. Llurba, G. Fretes, and R. Palau, “Classroom Emotion Monitoring Based on Image Processing,” Preprints, vol. 1, 2024. DOI: 10.20944/preprints202401.1323.v1.

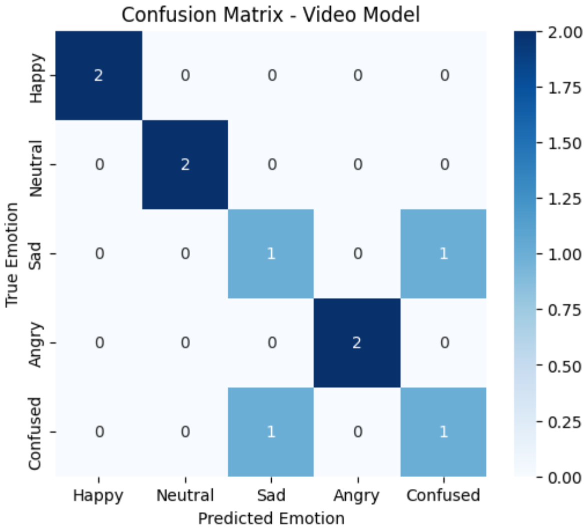

Fig. 5 Confusion Matrix: Video Model

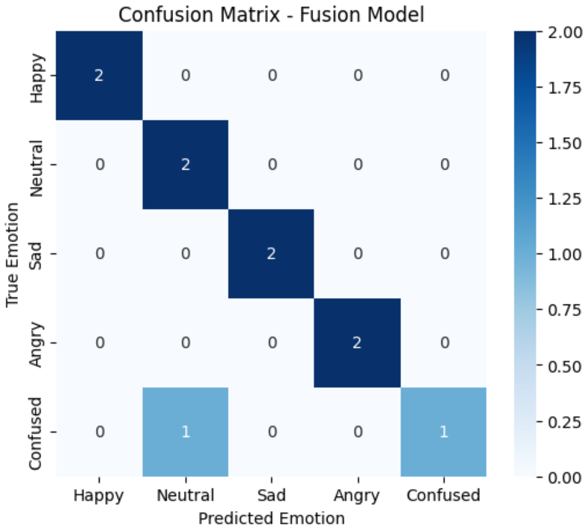

Fig. 6 Confusion Matrix: Fusion Model

[8] D. Olaniyan, M. O. Adebiyi, A. A. Adebiyi, and J. Olaniyan, “Enhancing Engagement in Virtual Classrooms: A Contactless Multi-Modal Emotion Detection Model,” in Proceedings of the SEB4SDG Conference, 2024, pp. 1–9. DOI: 10.1109/seb4sdg60871.2024.10630108.

[9] Y. He, X. Lu, D. Sun, T. Pan, Y. Qiu, and J. Liu, “Research on Teacher Classroom Teaching Speech Emotion Recognition Based on LSTM,” in Proceedings of the IEEE International Conference on Asian Language Processing, 2024, pp. 326–331. DOI: 10.1109/ialp63756.2024.10661152.

[10] S. Fakhar, J. Baber, S. U. Bazai, S. I. Marjan, M. Jasinski, E. Jasinska, M. U. Chaudhry, Z. Leonowicz, and S. Hussain, “Smart Classroom Monitoring Using Novel Real-Time Facial Expression Recognition System,” Applied Sciences, vol. 12, no. 23, 2022. DOI: 10.3390/app122312134.

[11] E. Alkhamali, A. Allinjawi, and R. Ashari, “Combining Transformer, CNN, and LSTM Architectures: A Novel Ensemble Learning Technique That Leverages Multi-Acoustic Features for Speech Emotion Recognition in Distance Education Classrooms,” Preprints, vol. 1, 2024. DOI: 10.20944/preprints202404.1456.v1.

[12] L. Shi, “The Integration of Advanced AI-Enabled Emotion Detection and Adaptive Learning Systems for Improved Emotional Regulation,” Journal of Educational Computing Research, vol. 32, 2024. DOI: 10.1177/07356331241296890.

[13] N. Siddiqui, M. A. Khalid, A. Ahmed, A. Aleem, M. Irfan, W. Ahmad, and F. B. Muslim, “An AI-Based Classroom Monitoring System Leveraging Computer Vision and Machine Learning,” in Proceedings of IBCAST, 2023, pp. 273–278. DOI: 10.1109/ibcast59916.2023.10713048.

[14] S. Worthington, “Automated Student Emotion Analysis During Online Classes Using Convolutional Neural Network,” in Proceedings of the Springer Series on E-Learning Advances, 2023, pp. 13–22. DOI: 10.1007/978-981-19-6525-8_2.

[15] L. Jie, Z. Xiaoyan, and Z. Zhaohui, “Speech Emotion Recognition of Teachers in Classroom Teaching,” in Proceedings of the Chinese Control and Decision Conference, 2020. DOI: 10.1109/CCDC49329.2020.9164823.

[16] G. Moise, E. Dragomir, D. Șchiopu, and L. Iancu, “Automated Emotion Recognition in Education Using Physiological Data and Deep Learning Models,” Journal of Advanced Intelligent Systems, vol. 17, 2024. DOI: 10.1007/s44196-024-00638-x.

[17] E. Yuan, “Research on Classroom Emotion Recognition Algorithm Based on Visual Emotion Classification,” Computational Intelligence and Neuroscience, vol. 2022, 2022. DOI: 10.1155/2022/6453499.

[18] S. Fakhar and M. Jasińska, “Novel Real-Time Facial Expression Recognition System for Classroom Monitoring,” Applied Sciences, vol. 12, no. 23, 2022. DOI: 10.3390/app122312134.

[19] R. Bojanic and A. Demark, “Facial Emotion Recognition and Educational Insights,” Advanced Computer Vision in Education, vol. 1, 2023. DOI: 10.20944/preprints202401.1456.v1.

[20] J. Yuan, “Classroom Teaching and Learning Analysis Using Convolutional Neural Networks,” Computational Analysis in Education Systems, vol. 2022, pp. 151–165, 2022. DOI: 10.1155/2022/6453499.