JURNAL INFOTEL
VOL. 18, NO. 1, FEBRUARY 2026, PP. 157–179. DOI:10.20895/INFOTEL.V18I1.1523

RESEARCH ARTICLE

Student Emotion Recognition from Low-Quality Videos Using Multimodal Deep Learning

Andi Mawadda Taiba[1,*], Rizki Yusliana Bakti[2], Muhammad Faisal[3], Muhammad Syafaat S. Kuba[4], Lukman Anas[5], Emil Agusalim H. T[6], and Fahrim I. Rahman[7]

[1,2,3,5,6,7]Department of Informatics, Universitas Muhammadiyah Makassar, Indonesia
[4]Department of Water Resources Engineering, Universitas Muhammadiyah Makassar, Indonesia

*Corresponding email: 105841103623@student.unismuh.ac.id

Received: June 17, 2025; Revised: November 24, 2025; Accepted: January 06, 2026.

Abstract: Emotion recognition plays a critical role in intelligent e-learning systems by enabling adaptive feedback and timely pedagogical interventions based on the affective states of students. However, most existing approaches rely heavily on visual facial cues, which are highly vulnerable to real-world conditions such as low-resolution video, partial facial occlusion, poor lighting, and unstable network connections commonly encountered in online learning environments. These limitations significantly degrade the performance of unimodal deep learning models. To address this challenge, this study proposes a multimodal deep learning framework for student emotion recognition that is robust to low-quality and occluded video input. The proposed model integrates visual and audio modalities through a hybrid architecture, combining a lightweight CNN-based visual feature extractor with a BiLSTM-based speech emotion model. An attention-based fusion mechanism is employed to adaptively weight cross-modal features, allowing the system to compensate for degraded or missing visual information using complementary acoustic cues. Experimental evaluations are conducted using publicly available datasets representative of realistic online learning scenarios, including DAiSEE and RAVDESS, with additional augmentation to simulate varying levels of occlusion and video degradation. The findings show that attention-based multimodal fusion improves the resilience and practicality of emotion-aware e-learning systems under non-ideal input conditions.

Keywords: student emotion recognition, multimodal deep learning, low-quality video, facial occlusion, e-learning systems

 ISSN: 2085-3688; e-ISSN: 2460-0997

1 Introduction

In fully online and hybrid higher-education environments, instructors have limited access to direct observational cues that traditionally support the interpretation of students' affective states. When learning interactions are mediated exclusively through digital platforms, emotional signals must be inferred indirectly from audiovisual information. Consequently, automatic emotion recognition has become an essential component of intelligent e-learning systems to support engagement monitoring and adaptive instructional feedback [1], [2]. Therefore, the recognition of emotions is a key component in the development of intelligent and responsive e-learning systems.

Current emotion recognition systems applied in educational contexts predominantly rely on visual information derived from facial expressions. Although deep learning architectures such as CNNs and Vision Transformers demonstrate strong performance under controlled acquisition settings, their effectiveness declines significantly in real online learning environments [3]. Factors including partial facial occlusion, low illumination, limited camera resolution, and unstable network conditions introduce substantial visual degradation, reducing the reliability of visual-only models [4], [5]. In addition, video compression artifacts and motion blur further exacerbate the reliability of visual-based unimodal approaches in consistently detecting emotions.

The limitations of unimodal visual-based approaches have motivated the development of multimodal emotion recognition frameworks that integrate complementary affective cues [6], [7]. Emotional expression is conveyed not only through facial appearance but also through vocal characteristics such as intonation, rhythm, and spectral patterns [8]. By combining visual and audio modalities, multimodal models are able to compensate for degraded or missing visual information and dynamically adjust modality contributions according to input quality [9], [10]. Although the multimodal approach shows promising potential, there are still a number of research gaps that need to be addressed. First, most multimodal models were evaluated using high-quality datasets collected under controlled conditions, thus underrepresenting real challenges in online learning, such as facial occlusion, low-resolution video, and network interference. Second, many architectures focus on improving accuracy without considering computational efficiency, making it difficult to implement in real-time on educational platforms with limited resources. Third, the problem of synchronization between modalities and signal quality imbalances where one modality dominates due to better data quality is still relatively underexplored in the context of online learning. This condition demonstrates the need for a multimodal approach that is not only accurate, but also resilient, lightweight and contextual.

Based on these problems, this study proposes and evaluates a multimodal deep learning framework for student emotion recognition that is specifically designed to be able to operate under low-quality video conditions and experience partial occlusion. The proposed model integrates two complementary modalities: visual and audio. Visual features are extracted using a CNN–Transformer hybrid architecture, while speech emotion dynamics are modeled using a Bidirectional Long Short-Term Memory (BiLSTM) network. The two streams are fused using an attention-based mechanism to improve robustness under occlusion and video degradation [11]. To simulate realistic online learning conditions, the training and evaluation process was carried out by applying various video quality degradation scenarios and audio interference through augmentation techniques oriented towards improving the resilience of the model. This study contributes both theoretically by advancing

robust multimodal fusion design under degraded input conditions and practically by providing a lightweight architecture suitable for deployment in real-world e-learning systems.

2 Literatur Review

2.1 Emotion Recognition in E-Learning

2.1.1 Importance of emotional feedback for engagement and adaptive learning

In online learning environments, students' affective states influence attention, persistence, and interaction throughout instructional activities. The absence of physical classroom interaction increases the risk that emotional disengagement remains undetected without computational support. As a result, emotion recognition techniques have become a key component in adaptive e-learning systems designed to improve engagement and learning effectiveness.

Recent studies show that the combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture with an attention mechanism can improve the accuracy of emotion recognition in video and audio content to surpass conventional methods [12]. This approach can be adapted in an online learning system to understand the emotional context of students even in low-quality or partially occluded video conditions. Positive emotions such as enthusiasm have been shown to increase knowledge retention, while negative emotions such as boredom can decrease learning performance. Thus, the application of multi-modal deep learning in e-learning has the potential to create a learning system that is more empathetic, responsive, and adaptive to students' emotional states, thereby enhancing a more personalized and effective learning experience.

2.1.2 Overview of datasets: FER2013, AffectNet, DAiSEE, EmoReact, etc

Positive and negative emotional states significantly influence student engagement and learning outcomes in digital environments. Empirical evidence indicates that positive affect, such as enthusiasm, enhances knowledge retention, whereas negative states such as boredom may reduce learning performance. These findings highlight the necessity of incorporating affect-aware mechanisms within intelligent e-learning systems to support adaptive and responsive instructional strategies. Several benchmark datasets have been developed to support research in emotion recognition under realistic conditions. AffectNet provides large-scale facial expression data captured under unconstrained environments, including variations in lighting and occlusion. DAiSEE contains annotated student engagement videos recorded in authentic e-learning scenarios, reflecting natural fluctuations in boredom, engagement, confusion, and frustration [12]. Meanwhile, EmoReact supports spontaneous multimodal emotion modeling under diverse environmental variations [10]. The availability of these datasets enables the development of robust deep learning models capable of handling degraded and partially occluded inputs.

Visual-based deep learning models remain the dominant approach for facial emotion recognition. Convolutional Neural Networks (CNN) effectively capture local spatial patterns such as eye and mouth movements, while Vision Transformers (ViT) provide global contextual modeling through self-attention mechanisms [13]. However, both architectures exhibit performance degradation under low-resolution and occluded video conditions. Re-

cent reviews suggest that integrating visual, vocal, and additional complementary modalities offers greater resilience in dynamic and uncontrolled real-world environments.

2.2 Visual-Based Deep Learning

2.2.1 CNNs and Vision Transformers (ViT) for facial emotion recognition

Convolutional Neural Networks (CNN) and Vision Transformer (ViT)-based models show strong performance in detecting facial expressions under ideal conditions, but both still experience significant decline in accuracy when faced with online learning videos with partial occlusion and low resolution. CNN excels at capturing local features such as mouth shape and eye movements, but tends to lose global spatial context, making it difficult to recognize emotions when part of the face is covered or visually degraded [4]. Meanwhile, ViT, which relies on a self-attention mechanism, is able to model the global relationship between parts of the face, but its performance is highly dependent on the resolution of the input image. In the low-resolution and low-light conditions common to online learning, ViT often results in inconsistent attention distribution and loss of focus on relevant areas of facial expression.

To overcome these limitations, recent research shows that the incorporation of local feature extraction and global context modeling through deep learning architectures is able to increase the resilience of emotion recognition systems to visual degradation and partial occlusion [14]. The findings are in line with a recent systematic review confirming that although Vision Transformer and the CNN-Transformer hybrid architecture show excellence in multimodal facial expression recognition, their performance is still heavily influenced by the visual quality of the input and the level of facial occlusion. The review also highlights that limited interpretability and sensitivity to visual degradation are major challenges in applying Transformer-based models to real-world scenarios, including online learning [15]. Although LSTM-based temporal modeling of a single modality is capable of capturing short-term emotional dynamics, the single-modal approach still has limitations in generalization and resilience to variations in visual context, making it less reliable in uncontrolled online learning environments.

2.2.2 Limitations under occlusion and low-resolution inputs

Convolutional Neural Network (CNN) and Vision Transformer (ViT)-based models have significant limitations in emotion recognition when faced with low-resolution video or partially occluded faces. CNN tends to only capture local features such as eye and mouth shapes, thus losing global spatial context when part of the face is invisible or exposed to lighting interference [**?**]. Meanwhile, ViT that utilizes the self-attention mechanism has an advantage in understanding the global relationships between facial areas, but its performance is highly dependent on the input resolution and image patch size. In poorly lit and low-resolution conditions, ViT often fails to maintain focus on key expression areas due to loss of texture and subtle color information.

Other limitations arise due to the compression and distortion of e-learning videos that lead to the loss of micro-expression details. A hybrid CNN–ViT model that combines local convolutions and global attention is able to fix some of these problems, but is still sensitive to occlusion and noise. To address these challenges, recent research encourages multimodal integration of deep learning by combining visual, audio, and physiological signals

so that the system can still recognize emotions even when visual data is compromised. This approach has been shown to improve model durability in online learning scenarios with low video quality and poor lighting [5]. Although several studies incorporate additional modalities such as physiological signals, the present study experimentally focuses only on visual and audio modalities to maintain consistency with the selected datasets and evaluation protocol.

2.3 Audio-Based Emotion Analysis

2.3.1 Speech emotion recognition using MFCC, spectrograms, BiLSTM, and CNN

Systematic reviews show that deep learning-based representation of acoustic and linguistic features, as well as precise fusion strategies, play an important role in improving speech emotion recognition performance [16]. Audio-Based Emotion Analysis focuses on processing sound signals to detect emotions through acoustic and prosodic features such as Mel-Frequency Cepstral Coefficients (MFCC), spectrograms, as well as deep learning models such as Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Network (CNN). Recent research by [8] showed that the combination of MFCC and deep neural network algorithms (DNN, CNN, and LSTM) can achieve up to 73% accuracy in speech-based emotion recognition in the RAVDESS dataset, after going through the stages of framing, windowing, normalization, and noise reduction to improve signal quality.

Meanwhile, a multi-modal approach that combines audio with visual and physiological cues is gaining popularity because it can increase the system's resistance to visual and acoustic interference [17] through the MiEmo platform shows that multi-modal feedback that combines music, color, and movement can strengthen the perception of emotions and increase user engagement in the context of therapy and social learning. These findings confirm that the use of optimized CNN and BiLSTM models on MFCC and spectrogram features remains a key approach in speech emotion recognition, especially when integrated with multimodal signals for real-time applications in education and healthcare settings.

2.3.2 Complementarity between visual and vocal emotion cues

The combination of visual and vocal cues plays an important role in improving the accuracy of the emotion recognition system because the two modalities complement each other. Visual cues, such as facial expressions, are able to describe emotions explicitly through muscle movements and mimics, while vocal cues reflect emotions through changes in tone, intensity, tempo, and prosodic contours of the voice. When video quality is low or faces are occluded, information from sound can be a major determinant in detecting emotional states [17]. Conversely, when audio signals contain noise or distortion, visual expressions can help clarify the true emotional context.

The synergy between these two modalities has been shown to improve the system's ability to recognize human emotions more naturally and consistently in a variety of environmental conditions. As research through the MiEmo platform shows that music and color-based feedback synchronized with facial and voice expressions results in stronger emotional responses, especially in learning and therapy contexts. This approach is in line with studied that confirmed that the combination of audio (MFCC, spectrogram) and visual (facial expression) features with multimodal deep learning architectures, such as CNN-BiLSTM or attention-based fusion [8], is able to improve the accuracy and resilience of the

system to occlusion conditions and low-quality inputs. Thus, the multimodal integration of visual-vocal is an important foundation in building a robust model of student emotion recognition in an e-learning environment.

2.4 Multi-Modal Fusion Strategies

2.4.1 Feature-level and decision-level fusion

In real-world emotion recognition systems, particularly within online learning environments, emotional cues are often distributed unevenly across different data streams. Visual information may be partially occluded, while audio signals may remain informative, or vice versa. This variability has motivated the use of multimodal fusion strategies to ensure that emotion recognition models remain functional under fluctuating input quality.

One approach to multimodal integration is feature-level fusion, in which representations from different modalities are combined prior to classification. In this setting, audio descriptors such as MFCCs and visual embeddings learned by convolutional neural networks are projected into a shared representation space. Early integration enables learning models to capture complementary relationships across modalities, which is beneficial for modeling complex emotional patterns [18]. However, this approach is sensitive to temporal misalignment and feature scale inconsistency, which may lead to modality dominance if not carefully handled. An alternative strategy is decision-level fusion, which defers integration until after each modality has produced an independent prediction. Instead of merging raw features, the outputs of modality-specific classifiers are aggregated using mechanisms such as confidence-weighted combination or probabilistic inference. This design offers greater robustness when one modality is unreliable or unavailable due to occlusion, noise, or transmission degradation [19]. Recent evidence indicates that combining these two strategies within a unified framework can mitigate their individual limitations. By jointly exploiting early cross-modal feature interactions and late-stage decision aggregation, hybrid fusion schemes have been shown to maintain stable emotion recognition performance in low-quality video conditions and dynamic online learning contexts.

2.4.2 Attention-Based fusion, late fusion, and tensor concatenation

In multimodal emotion recognition, not all input modalities contribute equally at every moment. The reliability of visual and audio signals may fluctuate due to occlusion, noise, or transmission artifacts, particularly in online learning environments. To address this variability, advanced fusion mechanisms have been introduced to dynamically regulate the contribution of each modality during the inference process. Attention-based fusion has emerged as a prominent solution for adaptive multimodal integration. Rather than treating all modalities uniformly, attention mechanisms assign context-dependent weights to modality-specific features, allowing the model to emphasize informative signals while suppressing unreliable inputs. Through self-attention and cross-attention operations, the fusion layer selectively captures salient emotional cues across modalities, improving robustness under conditions of partial data degradation [5]. In contrast, late fusion strategies operate at the output level by combining the predictions generated independently by each modality-specific model.

This approach avoids direct interaction between feature representations and instead relies on aggregation mechanisms to produce the final decision. While late fusion is gen-

erally less expressive than attention-based integration, it offers increased stability when one modality experiences severe quality loss, as each classifier remains decoupled from the others [18]. Tensor concatenation represents a complementary fusion technique in which features from multiple modalities are stacked into a higher-dimensional representation before classification. This approach preserves modality-specific information while enabling downstream networks to learn cross-modal correlations implicitly. However, without adaptive weighting, tensor concatenation may suffer from feature imbalance when one modality dominates the joint representation.

2.4.3 Challenges in synchronization and imbalance between modalities

The main challenge in multi-modal fusion systems lies in time synchronization and quality balance between modalities. Visual data typically has a different number of frames and sampling rates than audio data, so precise temporal alignment is needed so that emotional information does not lose context. In addition, a quality imbalance between visual and audio signals - for example, low-resolution video but clean audio can lead to bias in predictive outcomes [18]. To overcome this, methods such as dropout modality, dynamic weighting, and feature normalization are used so that the model can adjust the contribution weight of each modality adaptively during the training process.

2.5 Robustness Against Occlusion and Noise

2.5.1 Data augmentation techniques (masking, blurring)

To improve the model's resistance to interference such as closed faces or low image quality, data augmentation techniques are widely used. Methods such as masking, cropping, blurring, and random occlusion simulate real-life conditions where part of the face is covered by masks, hands, or shadows. With this strategy, the learning model recognizes emotional patterns even if the visual information is incomplete [20]. In addition, techniques such as specaugment on audio data which removes some of the time or frequency in the spectrogram can help audio-visual systems become more resistant to environmental noise and input distortion.

2.5.2 Generative models (GANs, Autoencoders) for reconstruction

Recent studies highlight that multimodal systems may integrate additional modalities such as physiological or textual signals to enrich affective representations. However, the present study experimentally focuses on visual and audio modalities to ensure methodological consistency, deployment feasibility, and alignment with video-based online learning scenarios.

3 Research Methodology

3.1 Research Framework

The framework integrates visual and audio modalities to capture complementary affective cues in online learning environments with degraded visual input. Each modality is pro-

cessed through the acquisition, pre-processing, and extraction stages of features using relevant deep learning models, then synchronized and combined through an attention-based fusion mechanism to produce a richer and more robust multimodal representation of noise and occlusion. Illustration shown in Figure 1.

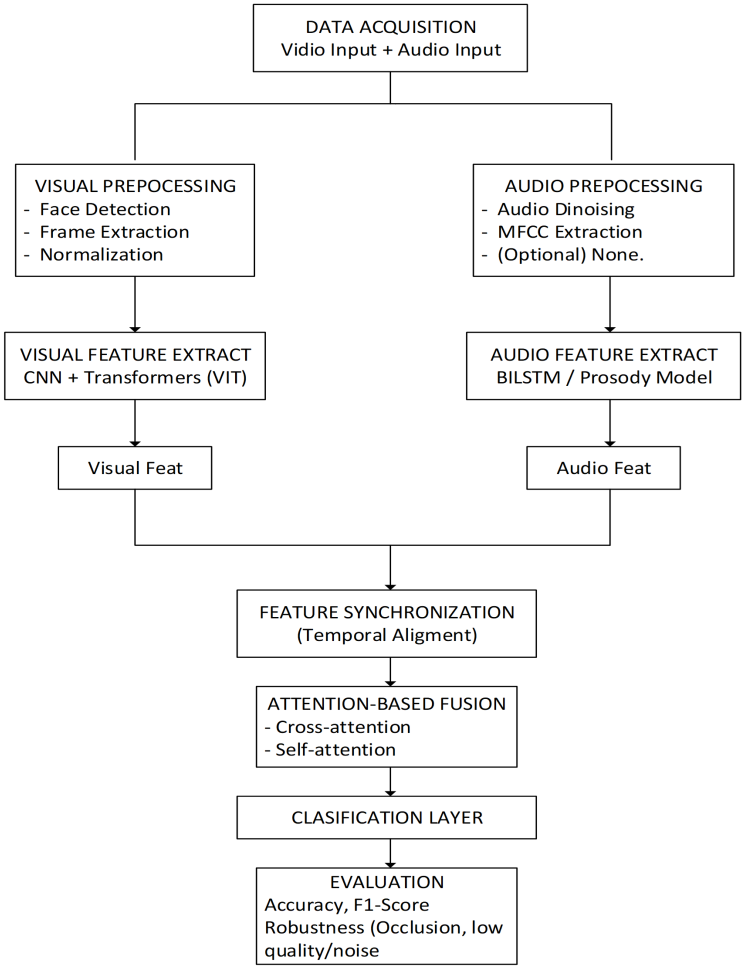

Figure 1: Proposed framework.

Figure 1 illustrates the complete architecture of the proposed multimodal emotion recognition system, which integrates two complementary modalities: visual and audio. At the data acquisition stage, facial video and speech audio are captured and preprocessed separately. In the visual stream, face detection, alignment, resizing, and normalization are applied before feature extraction using a CNN–Transformer hybrid architecture. This design enables the model to capture both local spatial features and global contextual dependencies from facial expressions. In the audio stream, raw speech signals undergo denoising, framing, and feature extraction using Mel-Frequency Cepstral Coefficients (MFCC). The temporal dynamics of speech are then modeled using a Bidirectional Long Short-Term

Memory (BiLSTM) network to capture sequential emotional patterns. The extracted visual and audio representations are temporally synchronized and combined through an attention-based fusion mechanism. This adaptive fusion layer dynamically adjusts modality contributions based on input quality, allowing the system to compensate for degraded visual information using complementary acoustic cues. The fused multimodal representation is subsequently passed to a classification layer to predict emotion categories. System performance is evaluated using accuracy, precision, recall, and F1-score to assess robustness under low-quality and occluded video conditions.

3.2 Dataset and Data Collection

These datasets were selected because their annotations capture affective states commonly observed during online learning activities, including engagement fluctuations under unconstrained recording conditions. The availability of video–audio data with varying visual quality, partial facial occlusion, and non-ideal lighting conditions enables realistic evaluation of emotion recognition models in practical e-learning scenarios [21], [22].

All data is processed by ensuring temporal synchronization between the video frame and the audio signal, as the timing alignment between modalities is an important factor in increasing the effectiveness of feature fusion and maintaining the emotional context [23]. In public datasets, this synchronization is available by default, while in self-recorded data, it is done through timestamp-based alignment.

To increase robustness to real-world conditions, various data augmentation techniques are applied, including random occlusion masking, Gaussian blur, resolution reduction, video compression, and noise addition to audio signals. This approach has been shown to help improve the generalization of the model against input quality degradation [10], [24]. Although the Transformer architecture exhibits competitive performance, the study adopts a lightweight CNN architecture to maintain computational efficiency and implementation feasibility in online learning systems.

3.3 Data Preprocessing

Data preprocessing was applied to visual and audio modalities. This pre-processing stage serves to ensure that each type of data is in optimal, standardized, and aligned condition before entering the feature extraction process. In visual data, the process includes face detection, face position alignment, and normalization of image size and intensity. In audio data, basic features such as MFCC and Zero-Crossing Rate are extracted after going through a process of signal cleaning and segmentation. The integration of all these stages results in clean, stable, and ready multimodal data that is ready for further processing at the feature extraction stage.

Figure 2 illustrates the data preprocessing workflow for the visual and audio modalities prior to feature extraction. In the visual pathway, the process begins with face detection to identify and crop facial regions from each video frame. The detected faces are then aligned to ensure spatial consistency across frames, followed by resizing and normalization to standardize pixel dimensions and intensity distributions. These steps ensure stable visual input suitable for deep learning-based feature extraction.

In the audio pathway, raw speech signals undergo denoising to reduce background interference, followed by framing and windowing to segment the signal into short-time

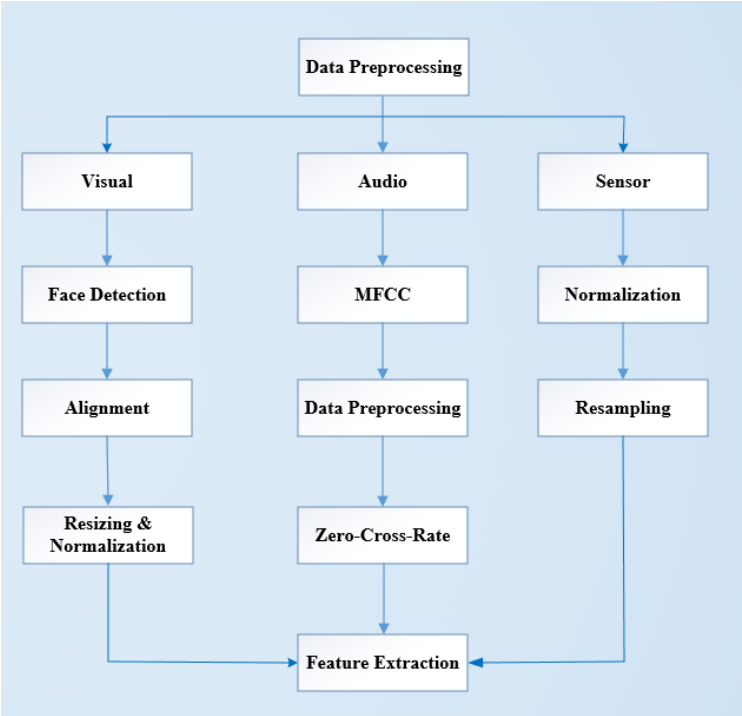

Figure 2: Flowchart.

intervals. Mel-Frequency Cepstral Coefficients (MFCC) are then extracted as the primary acoustic representation of emotional characteristics. Additional normalization is applied to stabilize feature distributions across samples. After preprocessing, both visual and audio streams are temporally synchronized to maintain alignment between facial expressions and vocal dynamics before entering the feature extraction stage.

3.4 Calibration, Validation, and Error Handling Strategies

Although the pre-processing stage is capable of reducing random noise and standardizing multimodal inputs, low-quality video conditions and partial occlusion still have the potential to result in systematic bias and representation instability. Therefore, a representation-level calibration mechanism is integrated before the fusion and inference process.

Calibration is carried out by first constructing a reference distribution from normal quality data to obtain mean parameters and variance as a stability baseline. In degraded data, the shift in embedding distribution to the baseline is calculated to detect variance inflation or average shifts. If the deviation exceeds the specified stability threshold during the training phase, scale adjustment and adaptive normalization are performed to reduce systematic distortion. This principle of systematic bias correction refers to the instrument calibration approach described by [25], but is adapted at the level of feature representation. To maintain consistency between modalities, the spatial alignment of visual and acoustic

features follows the concept of extrinsic calibration proposed by [26], so that the representation is in a calibrated space prior to the fusion process.

The reliability of the system under various input quality conditions is evaluated through a structured validation framework. The dataset is divided using stratified k-fold cross-validation by considering quality categories (normal, low-quality, and occluded) so that the degradation distribution is balanced in each fold. Evaluations were carried out separately for each degradation condition to identify patterns of performance degradation. A confusion matrix is used to calculate the sensitivity, specificity, and degree of misclassification, so that performance imbalances between classes can be analyzed in detail [27].

Since visual and acoustic degradation can occur with different degrees of severity, a reliability-based error handling mechanism is applied. The embedding variance of each modality is continuously monitored to detect representation instability. Each modality is then assigned a reliability score based on its distribution distance to the reference baseline. If the score is below the stability threshold, the contribution weight of the modality in the fusion process is adaptively reduced. This approach allows the system to experience a gradual degradation of performance (graceful degradation) instead of a sudden prediction collapse.

This approach is effective in moderate degradation scenarios where at least one modality still retains adequate semantic information. However, in conditions of total loss of one of the modalities or extreme occlusion, the effectiveness of weight adjustment becomes limited. In addition, distribution-based stability thresholds depend on the characteristics of the training data and may require readjustment on different domains.

3.5 Training Setup

The training process in this multimodal model is designed to ensure learning stability and resilience to varying data in the online learning environment. The dataset was divided into three subsets using stratified sampling, namely 70% for training, 15% for validation, and 15% for testing, so that the distribution of emotions classes remained balanced. The training was conducted with a batch size of 32 using the AdamW optimizer and an initial learning rate of 1×10^{-4}, accompanied by a Reduce-on-Plateau scheme to adjust the learning rate adaptively when validation performance is stagnant. The visual stream (CNN–Transformer) and the audio stream (BiLSTM) are processed in parallel before being combined in the fusion layer. The duration of the training ranged from 50–100 epochs, with an early stopping mechanism to prevent overfitting as well as dropouts (p = 0.3) on multiple layers to improve generalization.

In addition, the model was trained using the dropout modality technique, which is to randomly disable one of the modalities on a portion of the batch to increase the model's resistance to data loss due to occlusion or noise. This approach ensures that the model does not rely on a single modality when the input conditions are not ideal. Experiments were conducted on GPUs such as NVIDIA RTX-series to speed up computing in Transformer, CNN, and BiLSTM modules, and utilize mixed precision training to reduce memory usage and increase inference speed. With this configuration, the model is trained to be efficient, robust, and able to operate in real-time scenarios, as required by the emotion recognition system on modern e-learning platforms. Based on these considerations, this study adopts an attention-based fusion approach to integrate visual and audio information.

3.6 Evaluation Metrics

The Confusion Matrix is not only used as a basic evaluation tool, but also as a foundation for calculating a more comprehensive range of performance metrics. Emotion recognition in occluded and low-quality video conditions is highly susceptible to misclassification, so derivative metrics such as precision, recall, and F1-score are needed to assess the accuracy and sensitivity of the model on each class of emotions individually. This approach is important considering that some emotions have similar facial expressions when part of the face is closed or video quality degradation occurs.

The calculation of the Confusion Matrix's derivative metrics allows for a more in-depth analysis of the model's ability to distinguish complex emotions, as well as measure performance consistency in challenging visual scenarios. The following formulas are used to determine these values.

Measure the accuracy of predictions in a given class.

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k} \tag{1}$$

The model's ability to find all the correct data on a given class.

$$\text{Recall}_k = \frac{TP_k}{TP_k + FN_k} \tag{2}$$

Harmonic mean between precision and recall.

$$\text{F1}_k = 2 \cdot \frac{\text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k} \tag{3}$$

Overall Accuracy.

$$\text{Accuracy} = \frac{\sum_k TP_k}{\text{Total Samples}} \tag{4}$$

Macro Average.

$$\text{MacroF1} = \frac{1}{K} \sum_{k=1}^{K} F1_k \tag{5}$$

Weighted Average.

$$\text{WeightedF1} = \sum_{k=1}^{K} w_k \cdot F1_k \;\text{ and }\; w_k = \frac{N_k}{N} \tag{6}$$

Based on the formulas, each metric provides a different perspective on the model's performance in classifying students' emotions. Precision assesses the extent to which the model's predictions are reliable for each emotion class, while recall measures the model's ability to detect the entire correct sample under real conditions, including when some facial expressions are closed or visual quality is degraded.

The F1-score then acts as a measure of the balance between precision and recall, making it particularly relevant for this research scenario that tends to result in an uneven distribution of errors between emotion classes. In addition, the use of macro and weighted

averages allows for an assessment of the model's global performance on potentially unbalanced datasets. Thus, the evaluation based on the confusion matrix and its derivative metrics provides a more holistic picture of the robustness of the multi-modal model in dealing with the challenges of occlusion and low-quality video input that are often found in online learning environments. Based on the literature review that has been discussed, this study focuses on the experimental implementation on the integration of visual and audio modalities with attention-based fusion mechanisms. Other approaches discussed in this section are presented as broader conceptual developments in multimodal emotion recognition research. The experimental implementation of this study, however, is strictly limited to visual and audio modalities in accordance with the selected datasets and evaluation design.

4 Results and Discussion

The results of the study were analyzed through a gradual evaluation that included the performance of the visual model as a baseline, the analysis of the contribution of audio modalities through ablation studies, and robustness testing against visual occlusion to assess the resilience of the system in challenging online learning conditions.

4.1 Quantitative Results

This quantitative evaluation was conducted using the DAiSEE and datasets, which represent online learning conditions with variations in video quality and student engagement levels [22]. The dataset is designed to reflect realistic online learning scenarios, where the visual quality of video is often degraded due to device limitations, network conditions, and shooting angles.

Table 1: Quantitative performance of baseline models at different levels of visual occlusion using the DAiSEE dataset

Occlusion (%)	Accuracy (%)	Precision_Macro	Recall_Macro	F1_Macro	N
0	84.75	0.2368	0.2319	0.2338	800
20	86.00	0.2487	0.2540	0.2494	800
40	84.38	0.2369	0.2309	0.2332	800
60	80.25	0.2381	0.2262	0.2296	800

Table 2: Comparison of quantitative performance between the HOG-based baseline model and the CNN model proposed on the DAiSEE dataset

Model	Method	Accuracy (%)	F1_Macro
HOG + Logistic Regression	Manual Feature	84.75	0.2338
CNN (MobileNetV2)	(HOG) Deep Learning	93.88	0.2421

The performance of the baseline model based on the Histogram of Oriented Gradients (HOG) feature is presented in Table 1 using several evaluation metrics, namely accuracy,

precision macro, recall macro, and F1-macro. The test results showed that the baseline model was able to maintain relatively stable performance under conditions without occlusion to medium occlusion levels. However, when visual occlusion increases by up to 60%, the performance degradation becomes more significant, especially on F1-macro metrics. These findings show the limitations of manual feature-based approaches in capturing robust visual representations when some visual information is not available in its entirety.

To evaluate the potential for performance improvement, a comparison was made between the baseline model and the deep learning-based model. The results of the comparison are shown in Table 2 the MobileNetV2-based CNN model yields higher accuracy than the HOG approach, indicating the advantages of end-to-end representation in extracting relevant visual patterns from learning video data. Although F1-macro values are still affected by class distribution imbalances in the dataset, consistent improvements in accuracy suggest that deep learning approaches are more adaptive to visual quality variations than the classic manual feature-based method. The lower F1-macro value than accuracy is due to the class distribution imbalance in the DAiSEE dataset, where the Engagement class dominates the sample count. This condition is common in online learning data and causes accuracy metrics to appear high even though performance in minority classes is still limited. Confusion matrix-based evaluation allows for a more comprehensive analysis of the performance of multi-class classifications through metrics such as precision, recall, and F1-score [28], as recommended in comparative studies of classification models.

To further address the class imbalance problem identified in the DAiSEE dataset, an additional experiment was conducted using class-weighted cross-entropy loss during training. The class weights were computed inversely proportional to the frequency of each emotion category to reduce the dominance of majority classes and increase sensitivity toward minority classes.

The weighted model achieved an overall accuracy of 64.45% with an F1-macro score of 0.63. Although the overall accuracy is lower compared to the non-weighted configuration, the macro-averaged metric indicates a more balanced classification performance across emotion categories. A detailed class-wise analysis reveals substantial improvements in minority-class recall. The Confusion class achieved a recall of 0.94, while Frustration reached 0.99, indicating that the model became significantly more responsive to underrepresented emotional states. This improvement demonstrates that weighted learning effectively mitigates bias toward the dominant Engagement class.

The increase in recall for minority classes is accompanied by reduced precision in certain categories, particularly Frustration (0.35), suggesting a trade-off between sensitivity and specificity. This behavior reflects a controlled overcompensation effect commonly observed in imbalance mitigation strategies. The results confirm that class-weighted learning enhances fairness and robustness in multi-class student emotion recognition under real-world online learning conditions. This additional experiment directly addresses the class imbalance issue at the modeling level and strengthens the reliability of the proposed framework.

4.2 Visualization and Interpretability

The confusion matrix in Figure 3 is used to analyze the model's prediction distribution to each class of emotions. The results of the visualization showed that most of the test sam-

Table 3: Performance of visual model with class-weighted loss (DAiSEE dataset)

Class	Precision	Recall	F1-score
Boredom	0.77	0.63	0.69
Engagement	0.86	0.49	0.63
Confusion	0.54	0.94	0.69
Frustration	0.35	0.99	0.51
Macro Avg	0.63	0.76	0.63
Overall Accuracy: 64.45%			

ples were predicted to be Engagement classes, while Boredom, Confusion, and Frustration classes had relatively low prediction rates.

This pattern indicates an imbalance in the class distribution in the DAiSEE dataset that affects the model's ability to distinguish minority emotions, even though the overall accuracy obtained is relatively high. Confusion matrix analysis was used to evaluate the error distribution between classes of emotions and assess the predictive stability of multiclass models, as recommended in a comparative evaluation study based on confusion matrix [29].

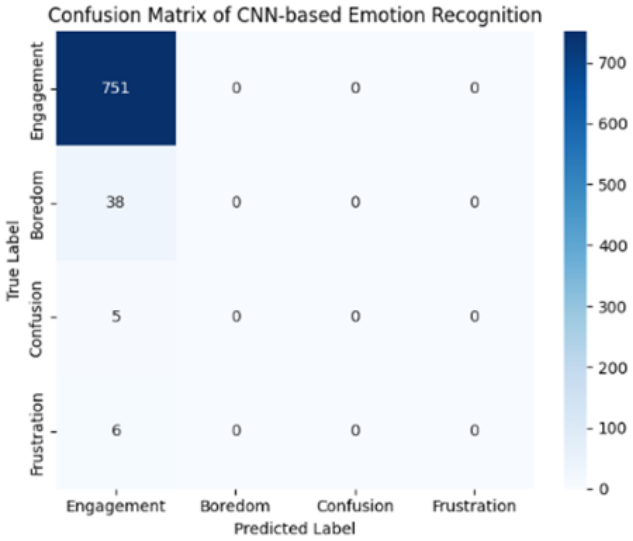

Figure 3: Confusion matrix results of testing MobileNetV2-based CNN model on the DAiSEE dataset.

To gain a deeper understanding of the model's decision-making mechanisms, Grad-CAM visualizations were performed as shown in Figure 4. The results of the Grad-CAM showed that the areas with the highest activation intensity were concentrated in the subjects' facial regions, specifically in the eyes and the area around the mouth. These findings suggest that the MobileNetV2-based CNN model utilizes semantically relevant visual regions in the emotion recognition process, rather than unrelated background or visual in-

Figure 4: Grad-CAM visualization showing the facial area as the dominant region influencing the emotion classification decision.

formation. These findings corroborate that the improvement in model accuracy is not only numerical, but also supported by semantically relevant decision-making mechanisms.

4.3 Multi-Modal Fusion Impact

In line with the scope of the methodology implemented in this study, the ablation analysis is focused on visual and audio modalities that are evaluated experimentally. The visual data used in this test is the result of preprocessing and augmentation through Roboflow, while the audio data is processed separately according to the methodologica flow described in the previous section. Other modalities presented in the conceptual framework, such as texts, are not included in the quantitative evaluation and are positioned as potential development in subsequent research.

To evaluate the contribution of each modality in the student emotion recognition system, an ablation study was conducted by comparing two main configurations, namely the visual-only model and the audio-only model. This approach aims to identify the relative role of each modality and understand its limitations when the system is faced with low-quality video input and experiences occlusion, as is common in online learning.

The test results showed that the visual-only model achieved an accuracy level of 61.58% with an F1-macro value of 0.60, while the audio-only model showed higher performance with an accuracy of 82.20% and an F1-macro value of 0.79. This difference indicates a significant variation in contribution between the two modalities in the process of emotion recognition.

Lower performance in visual-only configurations suggests that video quality degradation, facial occlusion, and lighting variations have a direct impact on the model's ability

Table 4: Ablation study

Variant	Accuracy (%)	F1_Macro
0	Visual - Only	0.615842
1	Audio - Only	0.822000

to extract representative visual features, especially for emotions with subtle facial expressions. In contrast, the audio modality shows better resilience because the emotional characteristics expressed through intonation, energy, and spectral patterns of sound are relatively unaffected by visual disturbances.

These findings confirm that visual and audio modalities provide emotional information that is complementary in nature, where visuals contribute to the representation of facial expressions and gestures, while audio captures emotional characteristics through acoustic signals. Thus, the results of this ablation study provide a strong empirical justification for the application of the multi-modal fusion approach to improve the reliability and resilience of students' emotion recognition systems in video input conditions that experience quality degradation and partial occlusion, in line with the findings of previous research. The results of this ablation study are in line with the findings of previous research which showed that the integration of information across modalities was able to improve the stability and accuracy of the emotion recognition system compared to the unimodal approach, especially in video input conditions that experienced quality degradation and partial occlusion [30].

4.4 Roboustness Analysis

The robustness analysis in this study was carried out using visual data that has gone through the stages of augmentation and video quality degradation simulation in Roboflow. This approach is designed to represent realistic online learning conditions, including the presence of facial occlusion, resolution drop, and video compression artifacts. Thus, the evaluation in this section directly measures the model's resilience in the face of low-quality input conditions as identified in the introduction.

To assess the resilience of the student emotion recognition model to the degradation of video input quality, robustness analysis was carried out by simulating different levels of visual occlusion. This test is designed to represent common online learning conditions, such as partially closed faces, suboptimal camera positions, and visual disturbances due to the environment. Evaluation was carried out by applying random occlusion to the input image area at 0%, 20%, 40%, and 60% levels, then measuring the performance of the visual-only model using accuracy and F1-macro metrics.

The test results showed a consistent decrease in performance as the occlusion rate increased. Under the condition without occlusion (0%), the model achieved an accuracy of 61.58% with an F1-macro value of 0.60. When the occlusion rate increased to 20%, the model's performance decreased to an accuracy of 56.78% and an F1-macro of 0.54. More significant decreases occurred at the occlusion rates of 40% and 60%, where the accuracy dropped to 49.09% and 42.27%, respectively, while the F1-macro values decreased to 0.45 and 0.35. This trend indicates that the model's ability to extract relevant visual features is increasingly limited when most facial information is covered.

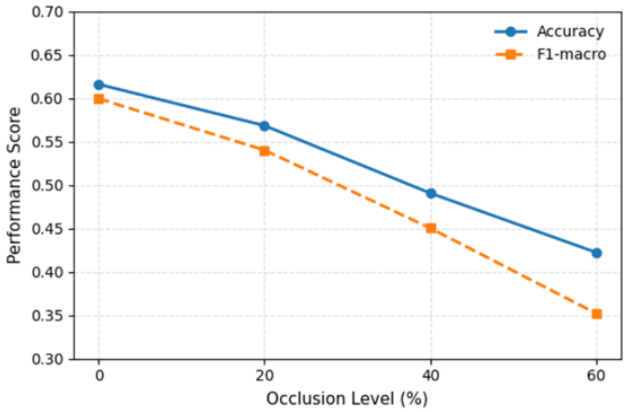

Figure 5: Robustness of the visual-only emotion recognition model under different occlusion levels, evaluated using accuracy and F1-macro metrics.

However, the visual-only model still shows gradual and stable performance degradation, rather than drastic or uncontrolled decline. This suggests that the representation of visual features that the model learns has a certain level of resistance to partial visual impairment. These findings confirm that although the visual modality has limitations in high occlusion conditions, this robustness test provides a strong empirical justification for the need for a multi-modal approach. By integrating audio information that is not affected by visual occlusion, the system is expected to be able to maintain students' emotion recognition performance more reliably in online learning scenarios and real-world environments. These findings strengthen the argument that the integration of audio and visual modalities is crucial to maintain the stability of the performance of students' emotion recognition systems under sub-ideal video input conditions.

The results of this test show that the visual-based approach has increasingly significant limitations as the occlusion rate increases. These findings, when associated with the results of the ablation study in subsection 4.3, provide an empirical justification that the integration of audio modalities has the potential to improve the performance stability of the emotion recognition system under degraded video input conditions.

The preprocessing pipeline plays a significant role in enhancing the robustness of the proposed multimodal framework. Face detection and alignment contribute to spatial consistency across frames, reducing feature instability under partial occlusion. Image normalization stabilizes intensity distributions, mitigating the impact of lighting variation and resolution degradation. In addition, data augmentation strategies such as random occlusion masking, Gaussian blur, and resolution reduction simulate real-world visual disturbances, enabling the model to generalize better under degraded input conditions. On the audio side, denoising and MFCC normalization improve the stability of temporal feature extraction, particularly in low-quality recording environments. These preprocessing steps collectively support the robustness behavior observed in the experimental results, especially under increasing occlusion levels.

4.5 Implementation Feasibility

To assess the feasibility of implementing the student emotion recognition system in a real learning environment, an implementation feasibility analysis was carried out by evaluating the inference latency and model processing speed in the scenario of a limited computing device (mobile/edge). This evaluation aims to ensure that the system not only achieves adequate classification performance, but is also capable of operating efficiently and responsively when applied to devices commonly used in the environment campuses, such as lecturer laptops, laboratory computers, and institutional edge servers.

Table 5: Quantitative performance of baseline models at different levels of visual occlusion using the DAiSEE dataset

Model	Batch Size	Avg_latency ms_per_batch	Std_latency ms_per_batch	Estimated_FPS samples_per_sc
Visual-only (MobileNetV2)	32	975.563562	20.406793	32.801553

The test results showed that the MobileNetV2-based visual-only model had an average inference latency of 975.56 ms per batch with a relatively small standard deviation, as well as an inference rate of about 32.80 samples per second. This value shows that the model is able to process visual data stably without significant latency fluctuations, thus supporting the use of the system in near-real-time student emotion monitoring scenarios.

These findings indicate a balanced trade-off between the level of accuracy and computational complexity. The selection of the MobileNetV2 architecture allows the system to maintain competitive classification performance while reducing computing load compared to more complex deep convolutional network architectures. Thus, this system is considered feasible to be implemented in the campus environment as a tool to monitor students' emotional state in a non-intrusive manner, especially in online or hybrid learning scenarios that require a fast and efficient system response. The findings of this study contribute to multimodal emotion recognition literature by demonstrating that spatial–temporal complementarity enhances robustness under degraded input conditions. The integration of CNN–Transformer visual modeling and BiLSTM-based audio dynamics confirms that cross-modal compensation improves resilience when one modality is compromised.

4.6 Theoretical Implications

The findings of this study contribute to the theoretical development of multimodal affective computing by demonstrating that robustness-oriented hybrid architectures can mitigate the limitations of unimodal emotion recognition models under degraded visual conditions. Specifically, the integration of CNN-based local feature extraction with Transformer-based global contextual modeling provides empirical evidence that combining spatial locality and long-range attention enhances resilience against occlusion and low-resolution inputs. Furthermore, the incorporation of BiLSTM for speech emotion modeling reinforces the theoretical understanding that temporal bidirectional dependencies are critical for capturing dynamic affective cues in audio signals.

4.7 Practical Implications

From a practical perspective, the proposed CNN-Transformer–BiLSTM hybrid framework offers a viable solution for deployment in real-world intelligent e-learning systems where video quality is often inconsistent and partial facial occlusion is common. The robustness experiments indicate that multimodal fusion preserves performance stability when visual information degrades, suggesting suitability for mobile-based learning platforms, low-bandwidth environments, and large-scale online classrooms.

5 Conclusion and Future Work

This study investigates student emotion recognition in online learning environments by proposing a multimodal deep learning approach designed to address the challenges of low-quality video input and partial facial occlusion. The experimental results demonstrate that integrating visual and audio modalities significantly improves recognition performance compared to unimodal approaches, particularly in conditions where visual information is degraded due to occlusion, low resolution, or poor lighting. The use of a lightweight CNN-based visual model combined with a BiLSTM-based audio model and attention-based fusion enables the system to exploit complementary emotional cues, resulting in more stable and reliable predictions across varying input conditions.

Ablation studies further confirm that the audio modality plays a crucial role in maintaining system performance when facial expressions are not fully observable, while visual cues remain valuable when video quality is sufficient. Robustness evaluations under different occlusion levels show a gradual and controlled performance degradation, indicating good generalization capability of the learned features. Moreover, inference latency analysis suggests that the proposed architecture achieves a balanced trade-off between accuracy and computational efficiency, making it feasible for near real-time implementation on resource-constrained educational platforms.

Despite these promising results, several directions remain open for future research. Future work may focus on deploying the proposed system in real-time learning management systems (LMS) to support continuous and non-intrusive monitoring of students' emotional states. In addition, cross-cultural and multilingual validation is required to ensure the generalizability of the model across diverse student populations. Further extensions could incorporate additional modalities, such as physiological signals or text-based emotion analysis from chat and forum interactions, to enrich emotional representations and enhance system robustness. Finally, exploring adaptive personalization strategies and long-term emotional trend analysis could further strengthen the role of emotion-aware systems in supporting effective and inclusive online learning.

References

[1] N. Ahmed, Z. Al Aghbari, and S. Girija, "A systematic survey on multimodal emotion recognition using learning algorithms," *Intelligent Systems with Applications*, vol. 17, p. 200171, 2023.

[2] M. Sajjad, F. U. M. Ullah, M. Ullah, G. Christodoulou, F. A. Cheikh, M. Hijji, K. Muhammad, and J. J. Rodrigues, "A comprehensive survey on deep facial expres-

sion recognition: challenges, applications, and future guidelines," *Alexandria Engineering Journal*, vol. 68, pp. 817–840, 2023.

[3] H. Gong, T. Luo, L. Ni, J. Li, J. Guo, T. Liu, R. Feng, Y. Mu, T. Hu, Y. Sun, *et al.*, "Research on facial recognition of sika deer based on vision transformer," *Ecological Informatics*, vol. 78, p. 102334, 2023.

[4] E. Boitel, A. Mohasseb, and E. Haig, "Mist: Multimodal emotion recognition using deberta for text, semi-cnn for speech, resnet-50 for facial, and 3d-cnn for motion analysis," *Expert Systems with Applications*, vol. 270, p. 126236, 2025.

[5] M. M. Islam, S. Nooruddin, F. Karray, and G. Muhammad, "Enhanced multimodal emotion recognition in healthcare analytics: A deep learning based model-level fusion approach," *Biomedical Signal Processing and Control*, vol. 94, p. 106241, 2024.

[6] S. Wu and D. M. Romano, "Robust emotion recognition using hybrid bayesian lstm based on laban movement analysis," *AI Open*, 2025.

[7] X. Qi, Y. Wen, P. Zhang, and H. Huang, "Mfgcn: Multimodal fusion graph convolutional network for speech emotion recognition," *Neurocomputing*, vol. 611, p. 128646, 2025.

[8] A. Chowanda, I. A. Iswanto, and E. W. Andangsari, "Exploring deep learning algorithm to model emotions recognition from speech," *Procedia Computer Science*, vol. 216, pp. 706–713, 2023.

[9] S. Woo, M. Zubair, S. Lim, and D. Kim, "Deep multimodal emotion recognition using modality-aware attention and proxy-based multimodal loss," *Internet of Things*, vol. 31, p. 101562, 2025.

[10] A. Wulamu, Y. Wu, X. Liu, Y. Zhang, J. Xu, and Y. Zhang, "Enhanced multi-modal emotion recognition using the feature level fusion," *Engineering Applications of Artificial Intelligence*, vol. 162, p. 112447, 2025.

[11] L. Cui, Y. Zhang, Y. Cui, B. Wang, and X. Sun, "A high speed inference architecture for multimodal emotion recognition based on sparse cross modal encoder," *Journal of King Saud University-Computer and Information Sciences*, vol. 36, no. 5, p. 102092, 2024.

[12] S. Guo, M. Wu, C. Zhang, and L. Zhong, "Emotion recognition in panoramic audio and video virtual reality based on deep learning and feature fusion," *Egyptian Informatics Journal*, vol. 30, p. 100697, 2025.

[13] G. Udahemuka, K. Djouani, and A. M. Kurien, "Multimodal emotion recognition using visual, vocal and physiological signals: a review," *Applied Sciences*, vol. 14, no. 17, p. 8071, 2024.

[14] J. Wang, H. Li, W. L. Woo, and S. Shan, "A single modality apparent first impression personality recognition model with temporal emotion based lstm," *Expert Systems with Applications*, vol. 259, p. 125114, 2025.

[15] I. Kus, C. Kocak, and A. Keles, "A systematic review of vision transformer and explainable ai advances in multimodal facial expression recognition," *Intelligent Systems with Applications*, p. 200615, 2025.

[16] A. Chaves-Villota, A. Jimenez-Martín, M. Jojoa-Acosta, A. Bahillo, and J. J. García-Domínguez, "Deep feature representations and fusion strategies for speech emotion recognition from acoustic and linguistic modalities: A systematic review," *Computer Speech & Language*, vol. 96, p. 101873, 2026.

[17] E. Chitti, R. Actis-Grosso, P. Ricciardelli, B. Olivari, C. Carenzi, M. Tedoldi, and N. A. Borghese, "Miemo: A multi-modal platform on emotion recognition for children with autism spectrum condition," *Computers in Human Behavior Reports*, vol. 17, p. 100549, 2025.

[18] R. Nirudeeswar, S. Thrishal, V. Shruthi, *et al.*, "Eranet: Emotion recognition based assistive learning network for autistic children," *Results in Engineering*, p. 106989, 2025.

[19] S. Gupta, P. Kumar, and R. Tekchandani, "An optimized deep convolutional neural network for adaptive learning using feature fusion in multimodal data," *Decision Analytics Journal*, vol. 8, p. 100277, 2023.

[20] S. Zheng, R. Wang, S. Zheng, L. Wang, and H. Jiang, "Adaptive density guided network with cnn and transformer for underwater fish counting," *Journal of king Saud university-computer and information sciences*, vol. 36, no. 6, p. 102088, 2024.

[21] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[22] A. Gordon, "Commonsense interpretation of triangle behavior," in *Proceedings of the aaai conference on artificial intelligence*, vol. 30, 2016.

[23] Y. Shang and T. Fu, "Multimodal fusion: A study on speech-text emotion recognition with the integration of deep learning," *Intelligent Systems with Applications*, vol. 24, p. 200436, 2024.

[24] J. Zhao, H. Zhu, and L. Niu, "Bitnet: A lightweight object detection network for real-time classroom behavior recognition with transformer and bi-directional pyramid network," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 8, p. 101670, 2023.

[25] J. Klueppel, M. Jurczak, U. Wallrabe, and L. M. Comella, "Calibration of multi-spectral photosynthetically active radiation sensor," *Available at SSRN 5711142*, 2025.

[26] J. Knuutinen, J. Backman, R. Linkolehto, and A. Visala, "Extrinsic parameter calibration methods of sensors present in a robot tractor," *Smart Agricultural Technology*, p. 101318, 2025.

[27] X.-L. Wu, J. B. Cole, A. Legarra, K. L. P. Gaddis, and J. W. Dürr, "Handling errors in the response: Considerations for leveraging unsupervised or incomplete data for genetic evaluations," *JDS communications*, 2025.

[28] D. Valero-Carreras, J. Alcaraz, and M. Landete, "Comparing two svm models through different metrics based on the confusion matrix," *Computers & Operations Research*, vol. 152, p. 106131, 2023.

[29] A. Theissler, M. Thomas, M. Burch, and F. Gerschner, "Confusionvis: Comparative evaluation and selection of multi-class classifiers based on confusion matrices," *Knowledge-Based Systems*, vol. 247, p. 108651, 2022.

[30] M. Faisal, T. K. Abd Rahman, D. Zainal, H. Mubarak, F. Shabir, N. Anwar, and I. Asrowardi, "Utilizing machine learning-based decision-making to align higher education curriculum with industry requirements," *International Journal of Modern Education and Computer Science*, vol. 17, no. 4, pp. 1–25, 2025.