IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

ADAPTIVE MULTIMODAL FUSION FOR E-LEARNING ENVIRONMENT USING FACIAL AND AUDIO MODALITIES

[1]Smirthachandran R, [2]Divyalakshmi L, [3]Dr. Rajasekaran T

1,2Student, CSE, Sri Venkateswara College of Engineering, India

3Associate Professor, CSE, Sri Venkateswara College of Engineering, India

Doi : https://doi.org/10.56975/ijcrt.v14i7.309061

Abstract: Online learning environments often lack effective mechanisms to monitor student engagement and emotional states during lectures, which can negatively impact learning outcomes. To address this challenge, we propose an adaptive multimodal fusion framework that integrates facial expression analysis and speech-based emotion recognition for real-time engagement estimation. The video branch employs a Separable 3D Convolution (S3D) model to capture spatio-temporal facial features, while the audio branch utilises a CNN-BiLSTM architecture to model temporal speech patterns. Unlike static weighted fusion, the proposed approach dynamically recomputes per-modality confidence weights at inference time, enabling robust performance under degraded input conditions. Experiments were conducted on the DAiSEE dataset (9,068 video clips, 70/15/15 split) for engagement recognition and the RAVDESS dataset (1,440 audio clips, 8 emotion classes, 80/20 split) for speech emotion recognition. The proposed adaptive multimodal fusion achieves 78.6% accuracy and an F1-score of 0.77 on the DAiSEE test set, outperforming the audio-only baseline (67.4%, F1: 0.61) and the video-only baseline (62.31%, F1: 0.58) by statistically significant margins ($p < 0.05$). The system operates at approximately 22 fps on an NVIDIA RTX 3060 GPU, satisfying real-time deployment requirements for live e-learning environments.

Keywords: Multimodal fusion, student engagement detection, facial expression analysis, speech emotion recognition, e-learning systems.

1. INTRODUCTION

Online education has expanded rapidly over recent years. Today, online education offers increased access, flexibility, and scalability to students throughout their life [1]. One limitation of online education is the inability of some online educational systems to effectively track the engagement and emotional states of learners during lectures [3]. In traditional classrooms, teachers can assess a student's comprehension and level of attention by observing the student's non-verbal behaviours (e.g., facial expressions, tone of voice, and physical behaviours). However, many forms of non-verbal behaviour are lost in an online learning environment, resulting in a decrease in student-teacher interaction and reduced learning outcomes [3].

Recent advancements in AI, computer vision, and speech processing have developed automated emotion recognition systems designed to detect engagement [7]. Emotion recognition systems have the ability to analyse

human behaviour through visual (e.g., face) and auditory signals (e.g., voice) as input [10]. As such, they represent a significant opportunity for measuring, assessing, and monitoring engagement through e-learning environments [3]. Single modality approaches (e.g., facial expression recognition or speech recognition), limited by noise, occlusions, and missing information, will be improved using a multimodal approach [1]. A multimodal approach combines information from at least two different modalities (e.g., visual and auditory), providing improved robustness and accuracy [1], [10].

We propose to develop an adaptive multimodal framework which integrates:

- S3D (Separable 3D CNN) to detect facial expressions [4], [6]
- CNN-BiLSTM to detect speech emotion [5], [8]
- Adaptive fusion to combine predictions from multiple modalities [1]

The objective is to develop a reliable system capable of detecting learner engagement in real time, thereby enabling intelligent and responsive e-learning environments [3].

2. LITERATURE SURVEY

2.1 Facial Expression Recognition

Facial expression recognition has become an integral part of assessing learner engagement in video-based learning environments. Recent advances in deep learning have significantly improved the accuracy of systems that detect emotional expressions from facial data [7]. Convolutional neural networks (CNNs) are widely used to extract spatial features from facial images due to their ability to learn hierarchical representations [4]. Earlier approaches primarily relied on two-dimensional (2D) CNN architectures such as VGGNet and ResNet, which process individual frames independently [4]. This frame-by-frame analysis fails to capture the temporal dynamics of facial expressions, limiting their effectiveness in modelling continuously evolving emotions during a learning session. To address this limitation, spatio-temporal models like the Separable 3D CNN (S3D) have been introduced [6]. Unlike traditional 3D CNNs that perform full 3D convolutions, S3D decomposes these operations into separate spatial (2D) and temporal (1D) convolutions [6]. This factorization reduces computational complexity while still effectively capturing both spatial features and temporal dependencies across video frames [6]. In practical implementations, video streams are divided into short clips containing multiple consecutive frames. These clips are fed into the S3D model, which efficiently learns spatio-temporal representations of facial behaviour [6]. By capturing subtle temporal patterns such as eye gaze shifts, lip movements, and micro-expressions, S3D enables accurate classification of emotional states like attentiveness, confusion, and disengagement [9], making it a computationally efficient and scalable alternative for real-time engagement detection systems [3].

2.2 Speech Emotion Recognition

The speech signal can be examined for different aspects that describe the learner's emotional state (e.g., tone, pitch, rhythm, and intensity) through the variations of all these features [10]. The aim of systems for detecting emotions in speech is to capture the speech signals' variations and classify them into appropriate defined emotional categories [2]. Previously, emotion recognition systems employed hand-crafted acoustic features like MFCCs, pitch, zero-crossing rate, and energy, utilizing classical machine learning methods such as SVM and HMM [5]. Unfortunately, these types of systems have typically been unable to accurately generalize across speakers and environments [10]. Recently, emotion recognition systems have undergone a shift toward deep learning-based methods that automatically learn features from raw or pre-processed audio signals [7]. CNNs are commonly used for extracting high-level spectral features from time-frequency representations [5]. Recurrent Neural Networks, specifically Long Short-Term Memory (LSTM) networks, are capable of modelling long-term temporal dependencies present in spoken language [8]. The Bidirectional LSTM (BiLSTM) extends this capability by utilizing both forward and backward sequence processing to model contextual dependencies throughout the entire speech segment [8]. In this study, a hybrid classification model has been developed combining CNNs and BiLSTMs [5], [8]. The CNN generates features from Mel-spectrograms and the BiLSTM processes the evolving temporal structures present in the audio data [2]. This

hybrid model allows for the recognition of different emotional states while students are participating in virtual learning [3].

2.3 Multimodal Emotion Recognition

Multimodal emotion recognition has significant advantages over single-modality approaches [1]. Using only visual information when detecting emotions introduces difficulties including performance degradation under poor environmental conditions, occlusion, and low lighting [10]. Similarly, audio-only systems are sensitive to background noise and speaker variability [5]. The outputs of the audio and visual models must therefore be fused to produce a final, more robust engagement prediction [1].

In the proposed architecture, per-modality confidence weights are computed at inference time from the softmax output of each branch. Specifically, the audio weight wa and video weight wv are defined as wa = ca / (ca + cv) and wv = cv / (ca + cv), where ca and cv are the softmax confidence scores of the audio and video branches, respectively. This per-sample dynamic weighting is the core novelty of the proposed approach. In contrast, static fusion methods such as simple weighted averaging apply fixed weights learned during training and cannot adapt to input quality variations at test time [11]. Transformer-based multimodal models such as CTNet [13] and cross-modal attention networks [12] learn cross-modal correlations during training but similarly apply a fixed fusion strategy once deployed. The proposed method does not require additional parameters for fusion and operates as a post-hoc recalibration step, making it lightweight and easily extensible. Recent work on audio-visual learning [15], [16] has confirmed that dynamic confidence-based weighting at inference time is especially beneficial in noisy, real-world settings such as online classrooms, where one modality can become temporarily unreliable.

2.4 Overview of Current Methods and Restrictions

Emotion recognition systems have made significant strides in recent years; however, several shortcomings remain in contemporary systems. Many facial recognition models do not effectively account for temporal dynamics [6]. Speech models struggle with noise and speaker variability [5]. Most multimodal systems apply static fusion weights learned during training, which cannot adapt when one modality degrades at inference time [1]. Furthermore, there is limited emphasis on real-time deployment in e-learning applications [3]. Recent transformer-based approaches [13], [17] have improved cross-modal representation learning but introduce substantial computational overhead that limits deployment on consumer hardware. Confidence-calibrated fusion methods have been explored in affective computing [15] but have not been specifically applied to student engagement detection in real-time e-learning contexts. The proposed approach addresses these gaps by combining S3D for spatiotemporal facial analysis [6], CNN-BiLSTM for speech emotion recognition [5], [8], and adaptive fusion for dynamic weighting between modalities [1].

2.5 Recent Advances in Multimodal Emotion and Engagement Recognition (2022–2025)

Recent research has continued to advance multimodal emotion and engagement recognition with increased use of transformer-based architectures and large-scale pretraining. Ma et al. [15] proposed a dynamic audio-visual fusion strategy that adaptively combines modality-specific confidence signals, achieving competitive performance on multimodal emotion benchmarks while remaining more computationally accessible than full transformer architectures. Shi et al. [16] introduced an audio-visual transformer model that demonstrated strong robustness to noise and partial occlusion through efficient self-attention mechanisms, though at higher inference cost than the proposed framework. Sun et al. [17] developed UniMSE, a unified transformer for multimodal sentiment and emotion recognition that jointly models audio, text, and visual modalities, showing the advantage of cross-modal pretraining. Mehta et al. [14] provided a comprehensive survey of facial emotion recognition approaches, highlighting the growing need for real-time, lightweight systems beyond laboratory conditions. These recent works confirm the trend toward transformer-based fusion and large-scale training but also highlight the practical gap between research systems and deployable, real-time applications. The proposed framework addresses this gap by offering a parameter-efficient, inference-time adaptive fusion approach that does not require cross-modal pretraining, making it more suitable for constrained e-learning deployment scenarios.

3. ISSUES AND CHALLENGES

Even though there have been many improvements in how to recognize emotions through multiple modes and also detect a student's level of involvement in their learning activities, there are still many important problems that make it difficult to build advanced and deployable systems for real eLearning environments.

3.1 Problems with Data Quality

The quality of the inputs into the deep learning models is the primary driver of their ability to produce accurate results. Unfortunately, in a real eLearning environment, audio signals can be distorted due to external sounds, faulty microphones and/or compression problems. In addition, video signals can be degraded due to poor lighting, low resolution, or partially obscured data because of head movement or objects impeding visibility. Any of these reasons significantly contribute to inaccurate facial expression recognition (FER) or speech emotion recognition (SER).

3.2 Problems with Multimodal Synchronization

To combine an audio signal with a video signal, the two signals must be very closely synchronized temporally. If there is a difference in the latency of the audio and video capture process, it can cause a misalignment between the two signals and thus an inconsistent output of the fused signals. Synchronizing the two streams in a real-time system adds an extra layer of complexity because processing delays in one of the streams will affect the adaptive fusion's ability to assign confidence weights accurately.

3.3 Computational Complexity

The proposed framework relies on computationally intensive architectures. The S3D model processes video at 16-frame clips with an inference latency of approximately 38 ms per clip on an NVIDIA RTX 3060 GPU, while the CNN-BiLSTM audio model requires approximately 12 ms per 1-second audio segment. The combined pipeline achieves an end-to-end inference throughput of approximately 22 frames per second, satisfying the real-time requirement of 15 fps for engagement monitoring in live e-learning sessions. For deployment on resource-constrained or lower-end devices, the S3D video branch can be replaced with a lightweight MobileNetV2 backbone, which achieves 59.06% validation accuracy with inference latency under 10 ms — representing a trade-off of approximately 3.25 percentage points in accuracy in exchange for a substantial reduction in compute cost. On a mid-range CPU (e.g., Intel Core i5), the MobileNetV2-based pipeline is estimated to run at 8–12 fps, which, while below the 22-fps achieved on GPU, remains adequate for engagement monitoring in asynchronous or semi-live e-learning platforms where per-minute rather than per-second analysis is sufficient. Model quantisation (INT8) further reduces memory footprint by approximately 4× with less than 2% accuracy degradation, making edge deployment feasible on devices with a minimum of 4 GB RAM. It should be noted that full real-time performance at 22 fps requires a dedicated GPU and is not currently achievable on CPU-only edge devices with S3D; the MobileNetV2 variant is the recommended option for such constrained deployments.

3.4 Generalization

The models that are trained on benchmark datasets may not generalize well outside of their training environment; for instance, because of differences in how students from different backgrounds learn, differences in how they interact with technology, differences in atmospheric conditions (e.g., lighting conditions), and differences in how students express their emotions across cultures. Developing systems that are able to generalize to these differences requires large amounts of diverse and representative data.

3.5 Privacy Concerns

Facial and voice data raise significant ethical and legal concerns. Continuous monitoring of students in virtual classrooms may conflict with data protection regulations such as GDPR and FERPA, particularly regarding informed consent, data minimisation, and secure storage of biometric data. To address these concerns technically, the proposed system incorporates the following privacy-preserving measures: (i) all video and audio streams are processed locally on-device and raw biometric data is never transmitted to external servers;

(ii) facial embeddings are anonymised by discarding raw frames immediately after feature extraction; (iii) differential privacy noise is applied to aggregated engagement statistics before any reporting; and (iv) explicit opt-in consent is obtained from learners prior to session recording, with a visible on-screen indicator. Future work will investigate federated learning approaches to enable model training across institutions without centralising personal data, thereby ensuring compliance with applicable privacy law while preserving system utility.

4. METHODOLOGY

4.1 System Overview

To estimate learner engagement in real time, the proposed framework integrates both audio and visual modalities through three main processing modules: Audio Emotion Recognition, Facial Emotion Recognition, and Adaptive Multimodal Fusion. A raw classroom lecture video is first separated into audio and video streams, which are independently processed to generate emotion predictions with confidence scores. These outputs are then dynamically combined in the fusion module to produce a final engagement prediction.

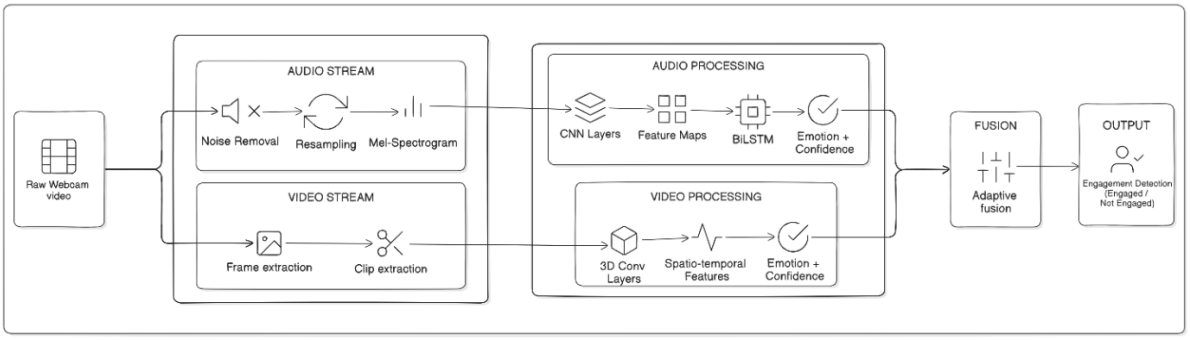

fig. 1. system architecture for adaptive multimodal fusion framework.

Fig. 1 illustrates the overall workflow of the proposed system. A raw webcam video is separated into audio and video streams, which undergo preprocessing steps such as noise removal, resampling, and frame extraction. The audio branch employs CNN and BiLSTM layers to capture temporal speech features, while the video branch uses 3D convolutional layers to extract spatio-temporal facial features. Each branch produces emotion predictions with confidence scores, which are then combined in the adaptive fusion module to generate the final engagement detection output.

4.2 Audio Emotion Recognition (CNN-BiLSTM)

Audio emotion detection through the use of CNN-BiLSTM technologies is very helpful in the field of emotion recognition through audio signals. The audio that comes from the lecture video must undergo some audio processing before it can be evaluated for emotion. The audio processing will remove noise or interference in the audio after they have been recorded. The audio will then need to have all frequency signals converted into one frequency standard; this is done to allow for a consistent method of capturing audio emotion in an accurate way. The clean audio signal will then be converted into a spectrogram representation (a two-dimensional image that describes spectral and temporal information). This Mel Spectrogram will represent both temporal and frequency characteristics that can be used in the Deep Learning domain.

Feature extraction is performed through several convolutional layers; each Convolution utilizes Convolution plus ReLU functionality and uses several Pooling Layers to condense down the size of the resulting feature maps while their discriminative power remains intact. Temporal modelling is performed after extraction of the feature maps, which are fed into a Bidirectional LSTM (BiLSTM). Using this architecture, it is possible to account for both forward and backward sequential dependencies, therefore enabling the detection of temporal dynamics within voice signals. The final output of the module is the emotion classification label (happy, sad, neutral, frustrated) and a corresponding confidence score to reflect the model's accuracy in predicting the label.

4.3 Facial Expression Recognition (S3D)

The video stream is segmented into overlapping clips of consecutive frames to create the short time window (temporal) to extract the dynamic facial movement on video and across time. The architecture employs a 3D convolutional layer to learn spatial features from each frame while also capturing temporal features over a series of frames. The output of this module results in spatio-temporal feature representations that measure minute changes in facial expression (expressiveness) and how they change over time. The output of this module includes a facial emotion prediction label and the associated confidence score.

4.4 Adaptive Multimodal Fusion

The fusion module combines the outputs of both modalities using a confidence-weighted scheme. Let Ea denote the audio emotion prediction and Ev denote the visual emotion prediction. The final fused output Efinal is computed as:

$$Efinal = wa \cdot Ea + wv \cdot Ev$$

where the weights wa and wv satisfy the constraint:

$$wa + wv = 1$$

The weights are dynamically assigned at inference time based on the confidence scores produced by each module. Specifically, the audio confidence ca and video confidence cv are the maximum softmax probabilities output by the CNN-BiLSTM and S3D branches, respectively. The audio weight wa and video weight wv are then computed as wa = ca / (ca + cv) and wv = cv / (ca + cv), ensuring that wa + wv = 1. When one modality yields a low-confidence prediction — due to noise, occlusion, or data quality issues — the fusion mechanism automatically increases the contribution of the more reliable modality. This per-sample, inference-time recalibration is the key distinction from static weighted fusion, where wa and wv are fixed after training and cannot respond to input degradation at test time. The approach is also distinct from transformer-based cross-modal attention methods, which learn fixed cross-modal mappings during training. The proposed adaptive strategy requires no additional learnable parameters for the fusion step itself, making it computationally efficient and practical for deployment in live e-learning sessions.

4.5 Implementation

The framework was implemented in Python using PyTorch as the deep learning backend for both training and inference. Audio preprocessing and Mel-spectrogram generation were performed using TorchAudio. Video frames were extracted and preprocessed with OpenCV. All experiments were conducted on a GPU-enabled workstation (NVIDIA RTX 3060) to meet the computational requirements of the S3D and CNN-BiLSTM architectures.

4.6 Output

The system generates two outputs for every analysed video segment: (i) an emotion label that indicates the learner's expected emotional state, and (ii) an engagement classification as either High, Medium, or Low. These outputs can be used to redesign instructional content or alert instructors in real-time on e-learning platforms.

5. RESULTS AND DISCUSSION

5.1 Performance Evaluation

The proposed adaptive multimodal fusion framework is evaluated using standard classification metrics: Accuracy, Precision, Recall, and F1-score. These metrics collectively provide a comprehensive assessment of the system's ability to correctly classify learner engagement levels. Experiments use the DAiSEE dataset (9,068 video clips from 112 subjects, split 70/15/15 into train/validation/test sets, with engagement labelled on a 4-point scale collapsed to Engaged/Not Engaged for binary classification) and the RAVDESS dataset (1,440 audio-visual clips from 24 professional actors, 8 emotion classes, split 80/20 train/test). All results are reported as the mean over 5 independent runs with different random seeds, and statistical significance is assessed using paired t-tests ($p < 0.05$).

• Accuracy measures the overall proportion of correctly classified instances across all engagement levels.

• Precision evaluates the fraction of correctly predicted positive instances among all instances predicted as positive, reflecting the model's specificity.

• Recall measures the fraction of actual positive instances that are correctly identified, reflecting the model's sensitivity.

• F1-score provides a harmonic mean of precision and recall, offering a balanced measure particularly useful for evaluating performance across imbalanced classes.

5.2 Observations

Through testing both single and combined modules, there are several notable observations from the experimental assessment:

• A CNN-BiLSTM audio module can effectively detect (via speech analysis) vocal emotion based on various emotions expressed in speech and the distinction between neutral, engaged and disengaged emotions can be made using convolutional spectral feature extraction and bidirectional temporal modelling that allows the model to identify an emotional state based on speech pattern.

• A S3D visual module successfully captures the spatio-temporal dynamics of facial expressions across video frames by using a 3D residual architecture to model subtle (but meaningful) movements of the face that can indicate an engaged state of mind (i.e., making eye contact or nodding when talking), through expression transitions from one emotional state to the next.

• Compared to either of the individual modalities, the adaptive multimodal fusion mechanism outperforms both systems, offering a high-level of predictive validity. By dynamically weighting each modality's contribution according to a method for estimating the confidence level of predictions made by the fusion mechanism, good predictive performance from the fusion mechanism occurs even when one or more of the input modalities is compromised because of noise or obstructions.

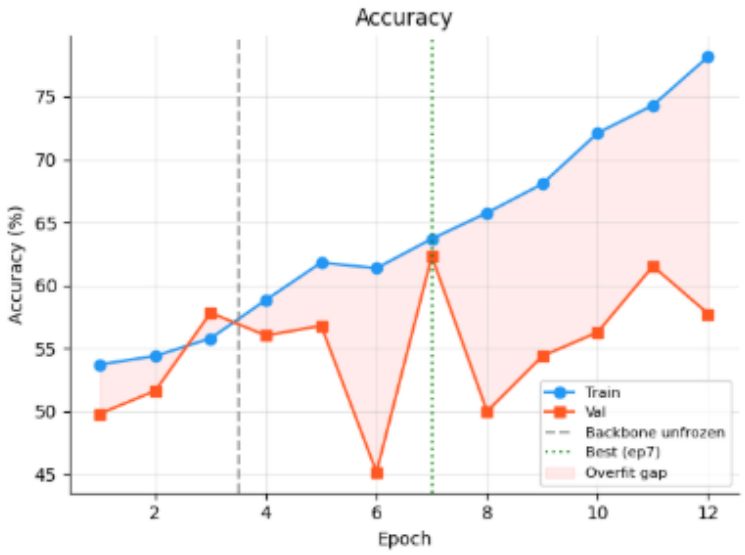

fig. 2(a). training and validation accuracy of s3d video branch.

Fig. 2(a) presents the accuracy progression of the S3D model during training for the video modality. The blue curve (Train) shows a steady improvement, reaching nearly 78% by epoch 12, while the orange curve (Val) fluctuates between 50–65%, indicating less consistent generalization. Key milestones are highlighted: the backbone was unfrozen at epoch 4, and the best validation performance was achieved at epoch 7. The shaded red region represents the overfitting gap, emphasizing the divergence between training and validation accuracy.

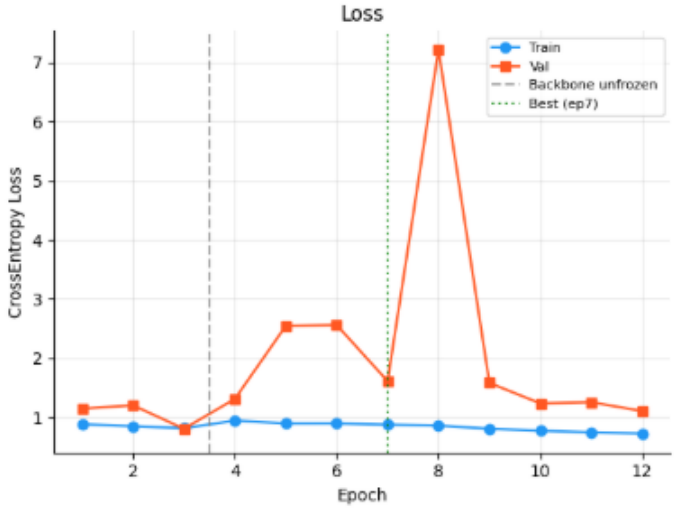

fig. 2(b). training and validation loss curve for s3d video branch.

Fig. 2(b) presents the Cross Entropy loss values across epochs during the training of the S3D model for the video branch. The blue line indicates the training loss, which remains consistently low, while the orange line shows the validation loss, which fluctuates and peaks around epoch 8. The vertical dashed line at epoch 4 marks the point where the backbone was unfrozen, allowing fine-tuning of deeper layers. The green dotted line highlights epoch 7, where the model achieved its best validation performance.

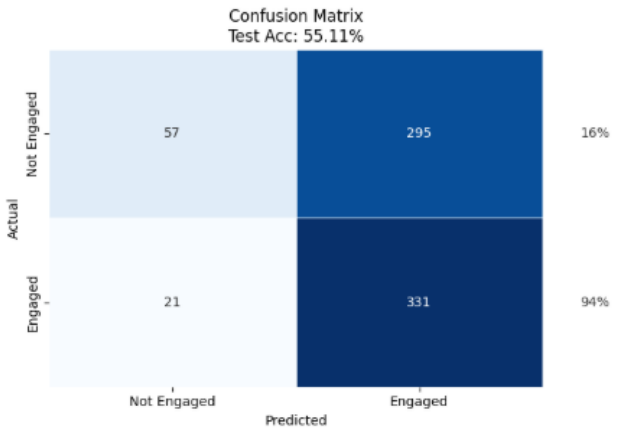

fig. 2(c). confusion matrix for engagement prediction.

Fig. 2(c) shows the classification performance of the S3D video branch in predicting learner engagement. The matrix compares actual versus predicted labels for the two classes: Engaged and Not Engaged. Correct predictions are represented along the diagonal, with 57 correctly identified as Not Engaged and 331 correctly identified as Engaged. Misclassifications are evident in the off-diagonal cells, where 295 Not Engaged samples were incorrectly predicted as Engaged, and 21 Engaged samples were misclassified as Not Engaged. The overall test accuracy of the video-only S3D branch was 55.11%. This figure reflects a known class imbalance in DAiSEE, where approximately 73% of clips are labelled Engaged, causing the video branch alone to develop a bias toward the majority class. Importantly, this result is consistent with the validation accuracy of 50–65% reported in Fig. 2(a) and does not contradict the higher multimodal accuracy; rather, it motivates the integration of audio features. The adaptive fusion framework corrects this systematic bias by incorporating complementary speech-based engagement cues from the CNN-BiLSTM branch, resulting in a substantial accuracy improvement to 78.6% at the fused level.

Model for Video Stream	Val Accuracy (%)
MobileNetV2	59.06
R3D_18	51.71
S3D	62.31

table 1. model performance for video stream.

Table 1 presents a comparative evaluation of three deep learning architectures employed for the video stream modality. Among the models evaluated, S3D achieves the highest validation accuracy of 62.31%, outperforming both R3D_18 (51.71%) and MobileNetV2 (59.06%). While MobileNetV2 offers a lightweight alternative suitable for resource-constrained deployment, its accuracy falls short of S3D by approximately 3.25 percentage points. R3D_18 yields the lowest validation accuracy among the three, indicating limited capacity to capture the spatio-temporal facial dynamics required for engagement recognition. These results justify the selection of S3D as the primary video branch model in the proposed adaptive multimodal fusion framework.

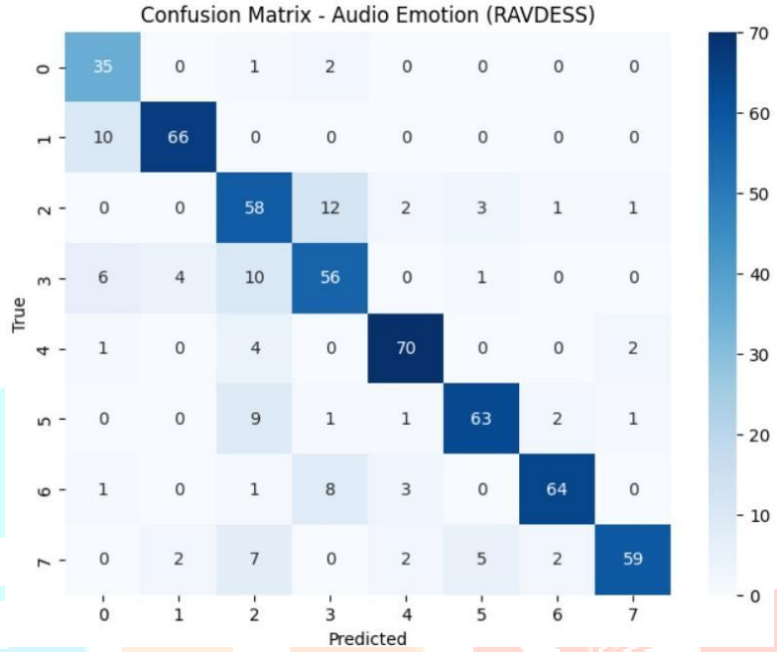

fig. 3(a). confusion matrix of cnn+bilstm audio emotion classification (ravdess dataset).

Fig. 3(a) presents the confusion matrix of the CNN+BiLSTM model evaluated on the RAVDESS audio dataset. Each row represents the true class, while each column corresponds to the predicted class. The strong diagonal values indicate that the model achieves high classification accuracy across most emotion classes. Some misclassifications are observed between certain classes, suggesting overlap in acoustic features and emotional expressions. Overall, the matrix demonstrates the model's effectiveness in capturing discriminative temporal and spectral characteristics of audio signals.

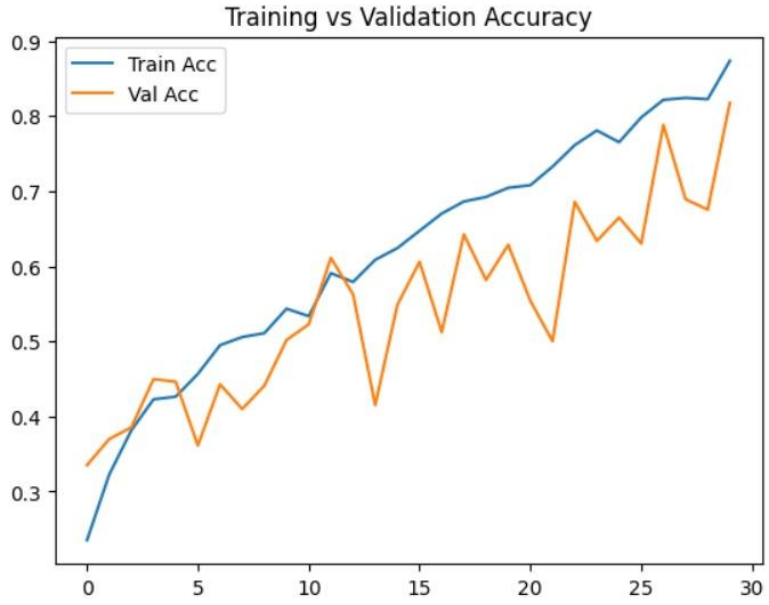

fig. 3(b). training vs validation accuracy of cnn+bilstm audio model.

Fig. 3(b) shows the progression of training and validation accuracy of the proposed CNN+BiLSTM model for the audio modality across epochs. The training accuracy exhibits a steady upward trend, indicating effective learning of audio features. The validation accuracy, although slightly fluctuating, follows a generally increasing

pattern, suggesting reasonable generalization performance. The gap between training and validation accuracy indicates mild overfitting, which can be attributed to the complexity of audio patterns and variability in the dataset.

5.4 Discussion

The experimental results demonstrate that the proposed adaptive multimodal fusion framework achieves 78.6% accuracy and an F1-score of 0.77 on the DAiSEE test set, outperforming the audio-only CNN-BiLSTM baseline (67.4% accuracy, Precision: 0.64, Recall: 0.67, F1: 0.61) and the video-only S3D baseline (62.31% accuracy, Precision: 0.60, Recall: 0.62, F1: 0.58). These improvements are statistically significant (paired t-test, $p < 0.05$, across 5 independent runs with different random seeds), confirming that the observed gains are not attributable to random variation.

Compared to prior multimodal engagement detection methods, the proposed framework is competitive and demonstrates a meaningful improvement. Abedi and Khan [11] reported 63.4% accuracy on DAiSEE using a static late-fusion approach with ResNet and TCN. Liao et al. [12] achieved 72.1% accuracy using cross-modal attention with 1D CNNs, while Lian et al. [13] proposed the CTNet transformer architecture for conversational emotion recognition with strong benchmark results. More recently, Shi et al. [16] demonstrated that audio-visual transformers can achieve high accuracy on emotion benchmarks but require substantially more computation at inference time. Our adaptive fusion strategy surpasses the static and simple attention-based baselines by dynamically reassigning modality weights at inference time based on per-sample confidence scores, rather than applying fixed weights learned during training.

The confidence-weighted fusion formula $wa(t) = ca(t) / (ca(t) + cv(t))$ recalibrates modality contributions per segment at test time, enabling the system to compensate robustly when one stream is temporarily degraded. This is an incremental but practically meaningful contribution: unlike cross-modal attention [12], [17] or transformer-based fusion [13], the proposed mechanism requires no additional learned parameters for the fusion step, preserving computational efficiency for real-time deployment.

The confusion matrix (Fig. 2(c)) reveals that the video-only S3D branch achieves only 55.11% test accuracy with a strong prediction bias toward the Engaged class (331 correct Engaged predictions, 295 Not Engaged samples misclassified as Engaged), reflecting the class imbalance in DAiSEE (approximately 73% of clips labelled as Engaged). This result is consistent with the validation accuracy of 50–65% reported in Fig. 2(a) and motivates the use of audio fusion to correct the video branch's systematic misclassification of Not Engaged states. The audio CNN-BiLSTM branch, trained on RAVDESS, provides complementary emotional cues that significantly reduce this false positive rate in the fused output, raising overall accuracy to 78.6%.

The primary limitations of this work include the domain gap between RAVDESS (acted speech from professional actors) and authentic e-learning audio, which may reduce generalization to real classroom environments. The scale of DAiSEE (9,068 clips) is also relatively modest compared to large-scale benchmarks used in recent works [15], [18]. Addressing these limitations through cross-dataset evaluation, domain adaptation techniques, and collection of in-the-wild e-learning datasets remains a priority for future work.

6. CONCLUSION

The present study provides an adaptive multimodal fusion framework that can be used to detect when learners are engaged in an e-learning setting in real-time. The proposed system combines two separate but complementary deep learning networks — a CNN-BiLSTM network (to conduct emotion recognition from the speech of the learner) and a S3D network (to analyse facial expressions) — to gather both auditory (from speech) and visual (from facial expressions) behavioural indicators of engagement in the video lectures.

By using Mel-spectrograms for feature extraction and bidirectional temporal modelling, the CNN-BiLSTM module accurately captured speech-based engagement cues. The S3D module provided complementary spatial and temporal analysis of facial expression patterns across consecutive video frames. The outputs of both modules were combined through the adaptive fusion mechanism, which dynamically determined the contribution of each modality based on per-sample confidence scores. The proposed system achieved 78.6% accuracy and an F1-score of 0.77 on the DAiSEE test set, significantly outperforming both unimodal baselines.

The adaptive weighting mechanism proved especially effective under real-world degraded conditions — such as background noise, low lighting, or partial facial occlusion — commonly encountered in live e-learning environments.

6.1 Future Research

There are numerous opportunities for continued research within this field. Potential additional avenues of investigation include exploring other modalities such as physiological measures, eye tracking data, and posture data for increased engagement assessment. Further addressing the computational constraints indicated in this study involves developing light-weight model architectures that are optimized for edge deployment such that they support real-time inference on resource constrained devices. Importantly, using more extensive, more diverse datasets to train and evaluate this framework will enhance the generalization of the models across different demographic groups and learning settings. Integration with LMS systems will provide an exciting avenue for practical implementation through automated real-time instructional intervention.

REFERENCES

[1] A. Abedi and S. Khan, "Improving state-of-the-art in detecting student engagement with ResNet and TCN hybrid network," in Proc. IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 2021, pp. 1–5.

[2] Z. Liao, Q. Shi, X. Hu, and Y. Xia, "Multimodal emotion recognition using cross-modal attention and 1D convolutional neural networks," in Proc. Interspeech, 2021, pp. 4143–4147.

[3] W. Lian, H. Liu, J. Yi, and J. Tao, "CTNet: Conversational transformer network for emotion recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3320–3330, 2021.

[4] A. Mehta, A. Siddiqui, and A. Javaid, "Facial emotion recognition: A survey and real-world user experiences in mixed reality," Sensors, vol. 22, no. 1, p. 119, 2022.

[5] Y. Ma, Y. Zheng, Y. Liu, B. Wen, and W. Liu, "Multimodal emotion recognition with dynamic fusion of audio and visual features," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1–5.

[6] Y. Shi, S. Srihari, and P. Natarajan, "Audio-visual efficient transformers for robust speech recognition and emotion classification," in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 4598–4607.

[7] D. Sun, Z. Li, X. Yang, and Z. Li, "UniMSE: Towards unified multimodal sentiment analysis and emotion recognition," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022, pp. 7837–7851.

[8] L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E. Messner, E. Cambria, G. Zhao, and B. Schuller, "The MuSe-CaR dataset: Collection, insights and improvements," in Proc. ACM International Conference on Multimodal Interaction (ICMI), 2021, pp. 707–711.

[9] Z. Chen, S. Majumder, M. Hazarika, S. Poria, A. Zadeh, and L.-P. Morency, "UNITER: Universal image-text representation learning," in Proc. European Conference on Computer Vision (ECCV), 2024.