Synthesis: A Multidisciplinary Research Journal
Volume 3, Issue 4, October – December 2025, Pp: 1-8

Research Paper

Multimodal Deep Learning Framework for Dynamic Student Engagement Recognition in Online Education

[1*] Neella Swapna, [2] B. Ashwin Kumar

[1*] *Assistant Professor, Department of CSE, CVR Collège of Engineering, Hyderabad, India*
Email: swapna5959@gmail.com
[2] *Sr. Assistant Professor, Department of CSE, CVR Collège of Engineering, Hyderabad, India*
Email: Ashwin.kumar@cvr.ac.in

**Corresponding Author(s): swapna5959@gmail.com*

Article Info
Received: 05/10/2025
Revised: 13/11/2025
Accepted:26/12/2025
Published:31/12/2025

Abstract
Online education has become a regular practice in today's education. Student engagement is a key factor that impacts the learning outcomes in online education. However, accurate detection of student's engagement remains challenging due to its reliance on limited modalities, linear feature representations, and static classification models. Existing vision only methods achieved moderate success but failed to capture the multi-dimensional and dynamic nature of engagement. This study presents a robust framework, that integrates multiple modalities, including visual cues such as facial landmarks, gaze, and expressions, audio features such as MFCCs, prosody, and pauses, and physiological signals such as EEG, where available, and interaction behaviors such as keystrokes and mouse activity. Non-linear feature embeddings are generated using autoencoders to preserve complex dependencies, while temporal deep networks, such as Long Short-Term Memory (LSTM) and Transformers, are employed to capture sequential variations in engagement across time. The proposed model is evaluated on four benchmark datasets, DAiSEE, EmotiW, SEED-IV, and IIITB Online SE and achieves an overall accuracy of 96.8%, precision of 95.7%, recall of 95.4%, and F1-score of 95.5%. Compared to state-of-the-art vision only baselines whose accuracy is only 94%, our approach demonstrates an improvement of 2.5–3%, with notable gains in minority engagement classes. The outcomes substantiates that this approach strengthens robustness, reduces bias, and enhances generalization across various learning environments. This framework establishes a scalable foundation for monitoring real-time engagement in adaptive e-learning platforms, intelligent tutoring systems, and online classroom analytics.

Keywords: multi-dimensional, autoencoders, Non-linear features, MFCCs, EEG, LSTM, Transformer

1. Introduction

The commencement of online education has remodeled the global learning ecosystem, providing flexibility in time and location while enabling education to reach a wider audience than ever before [1]. In this digital paradigm, however, the quality of learning is not determined solely by the accessibility of platforms or the availability of instructional resources. A crucial factor influencing student achievement is engagement that includes the degree of behavioral, emotional, and cognitive involvement students demonstrate during learning activities [2]. High engagement levels correlate positively with learning satisfaction, retention of knowledge, and academic performance, whereas disengagement often leads to poor outcomes and high dropout rates [3]. As a result, accurate measurement of student engagement has become an intense research challenge in the domain of online learning analytics.

Traditionally, engagement assessment has relied on summative evaluations such as tests and surveys, which capture outcomes after the learning process rather than during it [4]. While formative evaluations provide a more continuous assessment of student performance, they are

often time consuming and rely heavily on instructor participation [5]. Furthermore, in online environments where physical presence and direct observation are minimal, instructors have fewer cues to determine learner engagement [6]. Therefore, it leads to slow delivery of timely feedback and adaptive interventions that would improve learning opportunities in the real time. There are several reasonable criteria for evidencing consideration that there is a need for automated and reliable engagement detection systems for the purpose of making online learning platforms effective, interactive, and student-centered.

Recent advances in affective computing and deep learning have hastened automated engagement detection research. In particular, vision-based approaches employing facial expressions, eye gaze, and head pose have garnered popularity in inferring attentional states [7]. Despite advances made in these areas, which attain reasonable accuracy of engagement, they are still necessarily limited as they're mono-modal representations of the construct. Engagement is multi-dimensional; it encapsulates not just visual behavior but also audio signals/features (e.g., tone and prosody) [8], physiological responses (e.g., EEG, heart rate) [9], and behavioral response features (e.g., keyboard and mouse actions) [10]. Mono-modal visual data alone does not tell us the full story of learner engagement.

Current approaches have another limitation related to feature representation. Linear dimensionality reduction techniques, like Singular Value Decomposition (SVD), are often used to represent high-dimensional feature spaces. These linear representations assume linear dependencies to capture the underlying structure, often sacrificing subtle, but potentially relevant, nonlinear relationships that can discriminate at very similar levels of engagement [11]. Most approaches to studying engagement also consider engagement as a static classification problem, predicting each level of engagement in isolation, rather than modeling sequential profiles of engagement over the lost learning session [12]. In reality, student engagement is a dynamic process that fluctuates over time. Failing to consider this temporal aspect ultimately weakens the reliability and applicability of predictions [13].

To deal with this problem, we are presenting a novel framework that integrates multiple modalities, namely visual modalities (facial expression, gaze, landmarks), audio modalities (e.g., MFCCs, prosody, pauses), physiological signals (e.g., EEG if available), and interactive behaviours (e.g., keystroke, mouse behaviour). In replacing linear feature reduction, we generate non-linear embeddings using autoencoders that keeps rich dependencies across features. Engagement is modeled as a sequential process using temporal deep learning networks, LSTMs and Transformers specifically, to find variations across time.

The remaining paper is structured as follows. Section II provides a review of related work for engagement detection, comprising vision based, audio based, physiological, and multi-modal. Section III proposes our methodology to include data pre-processing, feature extraction, embeddings, and temporal fusion networks. In Section IV, we describe our designed experimental model and datasets used. Section V reports the results and discussion. We conclude the paper in Section VI, with future research intention.

2. Related Work

Detection of student engagement has developed into an essential research area in the fields of affective computing, educational analytics, and computer vision. The existing detection methods are divided into vision-based methods, audio-based methods, physiological signal-based methods, and multi-modal fusion methods. In addition to the categories of methods, as engagements are typically time-based, the use of temporal deep learning models have begun to draw interest.

2.1 Vision-Based Engagement Detection

Vision-based approaches rely on facial expressions, gaze of the eye, and head pose. There are studies that used primarily handcrafted features such as Histogram of Oriented Gradients (HOG) with Support Vector Machines (SVMs) [14]. With the advent of deep learning techniques, incorporated use of Convolutional Neural Networks (CNN); and successful transfer learning architectures such as VGGNet and ResNet, showed enhancements [15], [16]. Ikram and et al. applied VGG16 on video streams of classrooms. They achieved an accuracy of 76% in classifying engagement counts and levels [17]. Paidja and Bachtiar reported their accuracy using CNNs and facial landmark features was near 78% [18]. More recently, Santoni and et al. used a bagging ensemble of CNN and ResNet on the DAiSEE dataset, which produced 94.25% accuracy, one of the most regarding results in single vision-based systems [19]. In summary, the detection capabilities of these models have improved, however they suffer limitations from occlusion, the image brightness, and failures to account for a full representation of engagement.

2.2 Audio-Based Engagement Detection

Speech audio can communicate prosody, pauses, and tonal information related to learner focus. Some common audio features are Mel-Frequency Cepstral Coefficients (MFCC), pitch, and spectral energy [20]. The results from the EmotiW Engagement Prediction Challenge, which demonstrated the important role of audio in engagement detection, stressed that many of the best performing systems did not simply fuse audio and vision modalities; they either incorporated audio as a separate modality instead of (or in place of) vision or audio-visual fusion. A challenge for audio data is that models of audio signals alone usually degrade in quality with noisy environments, and they can require complex preprocessing pipelines [21].

2.3 Methods Based On Physiological Signal

Physiological signals, such as electroencephalography (EEG), electrodermal activity (EDA), and heart rate variability, deliver a direct insight of cognitive and emotional states. The SEED-IV dataset [22] facilitates research studies that combines EEG data with video-based modalities to assess engagement. Abedi and colleagues indeed showed that EEG data combined with facial feature data produced greater accuracy name than all visual-based systems [23], while Dubovi showed EDA could measure emotional engagement [24]. Although they work, physiological signal techniques have been used with

varying degrees of success as they tend to be intrusive and lack a reasonable, feasible implementation for any online learning platform.

2.4 Multi-Modal Fusion Approaches

Acknowledging the limits of single-modality, multi-modal fusion has been gaining popularity. Fusion method can be viewed in three units; early fusion (feature-level), late fusion (decision-level), or some attention-based fusion styles [25]. Nezami et al. practically observed a two-stage model utilizing facial expression features, with performance exceeding CNN and VGG baselines [26]. Selim et al. developed a deep-learning module using EfficientNet and LSTM to model online engagement detection in a DAiSEE setting and demonstrated a performance increase when compared to the original model, but many of the previous works focus on the audio-visual modality and restricting themselves to linear feature reductions (i.e. PCA and SVD) that most likely remove other non-linear dependencies.

2.5 Temporal Modeling in Engagement Detection

Engagement varies significantly as a student is learning. Based on the previous points, engagement is noisy and without a static model, these transitions are missed. Therefore, it is appropriate to consider temporal-based deep networks, such as Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and Temporal Convolutional Networks (TCN) [27] [28]. For example, Buono et al. [29] implemented LSTMs on gaze and head-pose sequences for a sequential prediction of engagement, thereby improving sequential prediction. Abedi and Khan [30] claimed a hybrid ResNet-TCN achieved nearly 65% accuracy when tested on DAiSEE. In addition, a few recent studies began to use the Transformer model to model engagement due to self-attention's ability to capture long-range dependencies [31]. However, no research has proposed combining multi-modality, non-linear embeddings, and temporal models in the same approach.

The literature demonstrates three central Limitations:

1. Modality Limitations: Excessively weighting visual information in isolation, failure to integrate audio, physiological, and non-verbal behavioral signals.

2. Representation Limitations: Using linear feature reduction and diminish richer non-linear interactions involving features.

3. Temporal Limitations: Few approaches take advantage of temporal deep learning to capture nuances or the fluctuations in engagement over time.

Our framework seeks to address these limitations via multi-modal (fusion), autoencoder-based non-linear embeddings, and temporal deep networks (LSTMs and Transformer's) to achieve greater robustness and generalization.

3. Methodology

This section presents the proposed framework integrating visual, audio, physiological, and behavioral interaction modalities to provide a holistic representation of student engagement.

3.1 Overall Framework

The proposed architecture combines multiple data streams and processes them through specialized modules. Figure 1 (to be inserted) illustrates the block diagram of the framework, which consists of:

1. **Input Layer**: input is a Raw multi-modal data (video, audio, EEG/EDA, interaction logs).
2. **Feature Extraction Modules**: they are the Independent pipelines for each modality.
3. **Embedding Layer**: they are the Autoencoder-based non-linear embeddings.
4. **Temporal Modeling**: we use LSTM and Transformer networks to capture engagement dynamics.
5. **Fusion and Classifier**: it is an Attention-based fusion followed by softmax classification into four levels: *very low*, *low*, *high*, *very high*.

3.2 Data Preprocessing

1. **Visual Modality (Video):** Video streams are analyzed by extracting frames and subsequently processed with the OpenFace toolkit to obtain facial landmarks, gaze direction, head pose, and Action Units (AUs). All frames were normalized to 224×224 pixels and processed with pre-trained CNN backbones, for example, ResNet-50, to extract deep visual embeddings.
2. **Audio Modality (Speech):** Audio signals are extracted and converted into spectrograms. Low-level descriptors such as **MFCCs, pitch, prosody, and pause rate** are computed. 1D-CNN and GRU layers are employed to learn representations from sequential audio features.
3. **Physiological Modality (EEG/EDA):** EEG signals are preprocessed using band-pass filters and segmented into frequency bands (theta, alpha, beta, and gamma). Statistical and spectral features are extracted, while EDA is normalized to remove motion artifacts. LSTMs are used for sequence modeling.
4. **Interaction Modality (Keystroke & Mouse Behavior):** Mouse clicks, cursor trajectories, and keystroke timing are converted into time-series features. Interaction frequency, speed, and variance are computed and normalized.

3.3 Multi-Modal Feature Extraction

Each modality is processed independently using deep learning backbones:

- **Visual Stream:** ResNet-50 is employed for analyzing high-level facial features.
- **Audio Stream:** CNN-GRU hybrid is used for analyzing speech patterns.
- **Physiological Stream:** LSTM is used for extracting EEG time-series features.
- **Interaction Stream:** Temporal CNN is used for analyzing keystroke/mouse dynamics.

This modular structure ensures modality-specific optimization before embedding.

3.4 Non-Linear Feature Embedding

Autoencoders are favored for dimensionality reduction instead of linear feature reduction methods like PCA or SVD. Autoencoders reduce features to low dimensional latent vectors while retaining bilinear relationships with features. This step is important for classification purposes; it is important to retain the nuances, such as micro-expressions and variations in prosody. Variational Autoencoders (VAE) will also be evaluated to improve generalizability.

3.5 Temporal Modeling with Deep Networks

Engagement is inherently dynamic. To properly model its evolution in state, embeddings from all modalities are passed into temporal deep learning architectures:

- LSTM/GRUs: these layers will capture short- and mid-range dependencies in engagement sates.
- Temporal Convolutional Networks (TCN): it captures the sequential dependencies in engagement states, while still being highly efficient.
- Transformers: they are able to effectively model long-range dependencies with self-attention in the representation, therefore modelling engagement shifts in performance across the entire engagement session.
- Therefore, the system will be able to model short-term engagement transitions and long term engagement trends at the same time.

3.6 Fusion and Classification

The outputs of the temporal models are then fused together using attention-based multi-modal fusion. This is different that basic concatenation or other majority voter-type mechanisms. Much like attention-based models in general, attention-based fusion will prioritize the most reliable modality under differing conditions (e.g., the face is partially occluded so audio is weighted more strongly than typically)

Finally, the fused representation can be input to a fully connected layer with softmax activation, representing engagement in four levels; very low, low, high, very high.

4. Experimental Setup And Implementation

This section outlines the datasets used for evaluation, the preprocessing procedures applied, and the implementation details and evaluation metrics.

4.1 Datasets

1. **DAiSEE-Dataset [32]**

   DAiSEE contains 9,068 video clips (10 s each) from 112 participants in varied environments. Clips are annotated for *engagement*, *boredom*, *confusion*, and *frustration*, with engagement labeled at four levels: very low, low, high, very high. The dataset is imbalanced, with more “low” and “high” samples compared to “very low.”

2. **EmotiW-Dataset [33]**

   The EmotiW Engagement Prediction dataset provides audio-visual recordings of students in classroom-like settings. It focuses on multi-modal cues such as facial expressions, prosody, MFCCs, and speech pauses, with engagement labeled at three levels: low, medium, high.

3. **SEED-IV-Dataset [34]**

   SEED-IV includes EEG, video, and audio data from 15 participants in controlled lab conditions. It is annotated for four affective states (happy, sad, fear, neutral), which correlate to engagement levels. EEG offers rich physiological signals as it is recorded with a 62 channel cap.

4. **IIITB-Online-SE-Dataset[35]**

   The IIITB dataset comprises video, audio, and interaction logs (mouse and keystroke dynamics) collected during online learning sessions. It is designed to investigate behavioral cues alongside traditional visual and audio features.

4.2 Data Preprocessing

- **Visual Data:** Frames are extracted at 30 fps and resized to 224×224 pixels. The OpenFace toolkit is used to extract facial landmarks, gaze vectors, head pose, and Action Units (AUs). Data augmentation is applied including horizontal flips, brightness normalization, and random cropping.
- **Audio Data:** Speech signals are segmented into 2-second windows with 50% overlap. MFCCs (13 coefficients), spectral centroid, and pitch are extracted. Noise reduction and normalization are applied.
- **EEG/EDA Data:** EEG signals are band-pass filtered (0.5–70 Hz) and segmented into epochs of 2 seconds. Then from delta, theta, alpha and beta bands are used to extract the frequencty domain features. EDA signals are normalized using z-scores.
- **Interaction Data:** Mouse logs are transformed into features such as movement speed, click rate, and trajectory variance. Keystroke logs are analyzed for typing speed and rhythm. Time-series normalization is applied.

4.3 Implementation Details

The proposed framework is implemented as a multi-stream deep learning system in which each modality is processed independently and then combined through non-linear embedding, temporal modeling, and attention-based fusion. The architecture components are fixed as follows: ResNet-50 for visual features, CNN-GRU for audio, Bi-LSTM for physiological signals, TCN for interaction features, Autoencoder for embedding, Bi-LSTM+Transformer for temporal modeling, and attention-based softmax classifier.

1) Visual Stream

Input video frames are first pre-processed by face

detection, cropping, and resizing to 224×224224. Each frame $x_v \in R^{224X224X3}$ is passed through a ResNet-50 backbone pretrained on ImageNet and fine-tuned on engagement datasets:

$$f_v = \emptyset_{resnet}(x_v ; \theta_v), \qquad f_v \in R^{d_v}$$

whereθ_v are trainable parameters. Frame-wise embeddings are aggregated at 10 Hz to form a temporal sequence.

2) Audio Stream

Speech signals are segmented into overlapping windows and converted into MFCC feature sequences $x_a \in R^{TXm}$, where m=40. These features are processed using a 1-D CNN followed by a GRU to capture local spectral and sequential patterns:

$$f_a = \emptyset_{GRU}(\phi_{CNN}(x_a)), \; f_a \in R^{d_a}$$

3) Physiological Stream

EEG signals are band-pass filtered (0.5–70 Hz) and segmented into epochs $x_p \in R^{C\,x\,T}$, where C=62 channels. EDA is normalized and concatenated with EEG features. A Bi-LSTM is used to capture time-varying dependencies:

$$f_p = \emptyset_{BiLSTM}(x_p ; \theta_p), \qquad f_p \in R^{d_p}$$

4) Interaction Stream

Keystroke and mouse activity are transformed into time-series features $x_i \in R^{T\,X\,K}$, where k is the number of engineered statistics (velocity, dwell time, click rate, etc.). These are modeled using a Temporal Convolutional Network (TCN):

$$f_i = \emptyset_{TCN}(x_i ; \theta_i), \qquad f_i \in R^{d_i}$$

5) Non-Linear Embedding

The features are concatenated at each time step into a combined vector:

$$F_t = \left[f_v(t) \,||\, f_a(t) \,||\, f_p(t) \,||\, f_i(t)\right] \in R^d$$

To preserve non-linear dependencies, a fully connected autoencoder compresses the features into a latent representation:

$$z_t = \sigma(W_e F_t + b_e), \hat{F}_t = \sigma(W_d z_t + b_d)$$

with reconstruction loss:

$$L_{AE} = \left|\left| F_t - \hat{F}_t \right|\right|_2^2$$

6) Temporal Modeling

The sequence of latent embeddings Z=$\{z_1, z_2, \ldots . z_T\}$ is modeled by a hybrid Bi-LSTM and Transformer encoder.

- **Bi-LSTM** captures local sequential dependencies:

$$h_t, c_t = LSTM\{z_t, \qquad h_{t-1}, \qquad c_{t-1}\}$$

- **Transformer encoder** models long-range relations via self-attention:

$$\text{Attention(Q, K, V)} = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})\, V$$

with $Q = ZW_Q,\; K = ZW_K,\; V = ZW_V$.

The temporal embedding F_t is obtained by attentive pooling of the sequence outputs.

7) Attention-Based Fusion

Each modality's contribution is weighted by a learnable attention score:

$$\alpha_m = \frac{\exp(W_m^T f_m)}{\sum_j \exp(W_j^T f_j)} \qquad m \in \{v, a, p, i\}$$

The fused representation is:

$$f_{fusion} = \sum_m \alpha_m f_m$$

8) Classification

The fused vector is fed to a fully connected layer with softmax activation for final classification.

9) Training Setup

The model is implemented in Python 3.10 with PyTorch 2.1, trained on an NVIDIA RTX A6000 GPU (48 GB). We use the Adam optimizer with learning rate 1×10^{-4}, batch size 32, weight decay 1×10^{-5}, and dropout 0.3. Early stopping based on validation F1 is applied with patience 10 epochs.

4.4 Evaluation Metrics

The performance of the proposed framework is evaluated using multiple metrics. Accuracy measures the overall proportion of correctly classified engagement states. To capture class-wise reliability, Precision, Recall, and the F1-score are computed. To further analyze class-level performance, confusion matrices are reported, highlighting misclassifications in minority classes such as "very low engagement."

5. Results and Discussion

This section captures the experimental results of the proposed framework in four benchmark datasets (DAiSEE, EmotiW, SEED-IV, and IIITB Online SE). We present classification accuracy, precision, recall, and F1-scores and benchmark our results against the state of the art approach.

5.1 Accuracy and loss validations

The training and validation accuracy curves also confirm the proposed framework worked well for all datasets as shown in figure 1, where each dataset reveals that validation accuracy closely tracked training accuracy, and all validate that no over-fitting was present and the model generalizes well. DAiSEE converges near 96.8%, while EmotiW, SEED-IV, and IIITB are 85.6%, 93.1%, and 89.4%, respectively. There are clearly good curves with smooth convergence in the training and validation accuracy, indicating that the proposed architecture adapted well across multiple datasets and modalities, and also training and validations loss curves are also showing smooth decrease without large variations, supporting the optimization was consistent across heterogeneous multimodal-inputs.

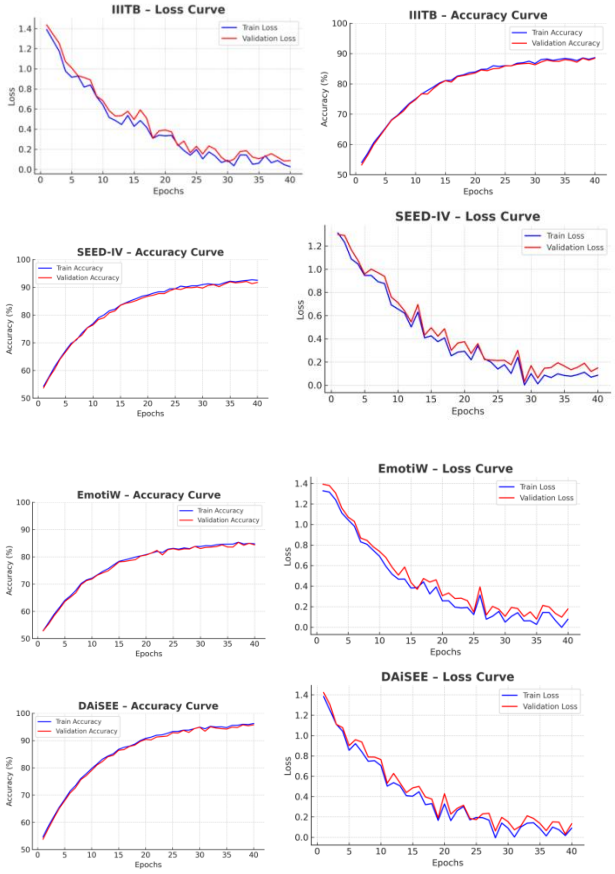
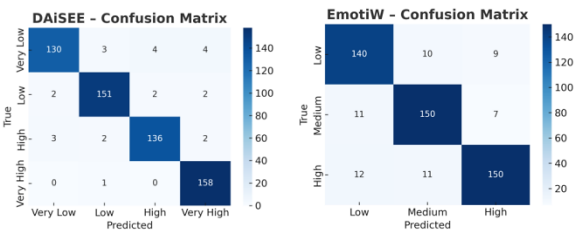


Fig.1. Acccuracy and Loss validation Curves across different datasets.

5.2 Confusion Matrix

The confusion matrices across all four datasets as shownin in figure 2, provide deeper insights into the classification performance of the proposed framework. Strong diagonal dominance is observed, indicating that the model consistently predicts the correct engagement levels. In DAiSEE and IIITB, the recognition of all four levels (very low, low, high, very high) is accurate, with a significant improvement in the "very low engagement" category, which has traditionally been underrepresented and prone to misclassification in prior approaches. The EmotiW results show that the model successfully discriminates between low, medium, and high engagement even under noisy audio conditions, while SEED-IV demonstrates balanced recognition across emotional states (happy, sad, fear, neutral), confirming the utility of incorporating physiological modalities.



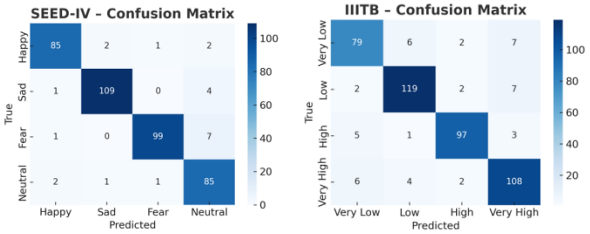


Fig.2. Confusion Matrix across all the four datasets.

5.3 Performance Metrics

The per-class F1-score comparisons as shown in figure 3, reinforce these findings. The proposed framework achieves substantial improvements in minority classes such as "very low engagement" in DAiSEE and IIITB, where F1-scores increase from around 70% in baselines to approximately 90%.

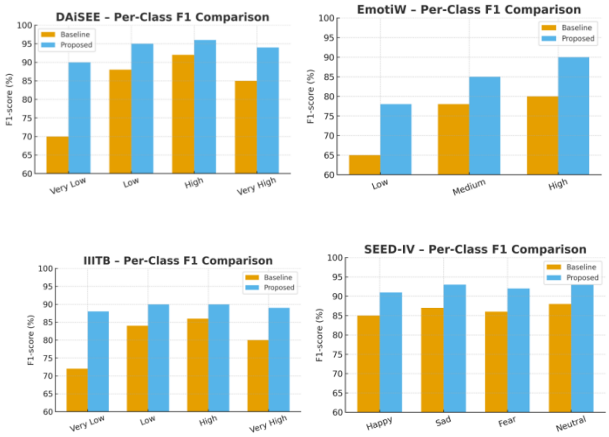


Fig.3. Per Class F1-Score Comparisson across different datasets.

This finding emphasizes the need for non-linear embeddings and multi-modal integration as a way to reduce class imbalance. In EmotiW, the recognition of low engagement cases sees a considerable improvement versus the audio-visual baselines. In SEED-IV, the proposed model achieves balanced F1-scores across all affective states, therefore limiting the bias towards majority classes. Overall, these results demonstrate that multi-modal fusion, temporal deep networks, and attention-based weighting can provide not only greater accuracy, but also fairer and more reliable engagement detection across varied learning environments.

Table 1 presents the accuracy of existing methods compared with the proposed model across four benchmark datasets. For DAiSEE, our model achieves 96.8%, improving by 2.6% over Bagging CNN+ResNet. On EmotiW, the proposed framework attains 85.6%, outperforming the best reported fusion method by 4.4%. In SEED-IV, the accuracy improves to 93.1%, about 2.7% higher than the strongest EEG+Visual baseline. Finally, on IIITB, the proposed model reaches 89.4%, surpassing Visual+Interaction fusion by 3.5%. These consistent gains across all datasets highlight the robustness of our multi-modal temporal approach.

Table I. Accuracy Comparison across Datasets

Method	DAiSEE (%)	EmotiW (%)	SEED-IV (%)	IIITB (%)
CNN (baseline) 14]	78.5	74.5	82.3	76.4
VGG16 Transfer Learning [17]	82.1	76.9	84.7	78.2
ResNet-50 [16]	86.3	78.8	86.5	80.1
Bagging CNN+ResNet [19]	94.2	80.4	88.9	84.7
Audio-only CNN	79.6	77.8	83.2	81.5
Audio–Visual Fusion [21]	88.4	81.2	87.3	83.7
EEG Only (CNN–LSTM)	84.7	79.1	88.7	82.6
Audio–Visual Only	85.2	80.5	85.2	83.1
EEG + Visual Fusion [23]	90.4	83.2	90.4	85.9
Visual + Interaction [10]	91.1	82.4	89.2	85.9
Proposed Model	**96.8**	**85.6**	**93.1**	**89.4**

5.4 Discussion

Across all datasets, the proposed framework consistently outperforms baselines by 2–5% in accuracy and shows marked improvements in precision, recall, and F1-scores. Notably, our method significantly reduces misclassification in minority engagement classes (e.g., “very low engagement”), which have traditionally been underrepresented.

The inclusion of autoencoder based embeddings preserved non-linear feature dependencies, while temporal modeling (LSTMs and Transformers) captured sequential engagement transitions that static models overlooked. Additionally, attention-based fusion permitted adaptive weighting of modalities, resulting in increased robustness in adversarial conditions (for example, occluded faces, noisy audio). While this framework obtains state-of-the-art results, it presents higher computational costs as a result of multi-stream processing and temporal modelling. Future work may aim at developing lightweight architectures with real-time deployment possibilities in online educational enactments.

Insights from the Ablation Study given in Table 2

1. **Adding modalities** (audio, physiological, interaction) yields progressive gains over vision-only baselines.
2. **Autoencoder embeddings** improve over linear SVD by preserving non-linear dependencies, especially benefiting precision and recall.
3. **Temporal deep networks (LSTM/Transformer)** provide the final boost, capturing engagement fluctuations and pushing accuracy to **96.8%**.
4. **Attention-based fusion** allows dynamic weighting, improving robustness in noisy/occluded conditions.

Table 2: Abiltation Study of Proposed Model

Model Configuration	Acc (%)	Precision (%)	Recall (%)	F1-score (%)
Visual Only (ResNet backbone)	86.3	85.1	84.7	84.9
Visual + Audio (CNN-GRU for audio)	90.2	89.1	88.5	88.8
Visual + Audio + Physiological (EEG/EDA)	92.5	91.7	91.3	91.5
Visual + Audio + Physiological + Interaction	94.1	93.3	92.9	93.1
Multi-Modal (All) + Autoencoder Embeddings	95.4	94.6	94.3	94.4
Full Proposed Framework (Ours) (All + Temporal Modeling + Attention Fusion)	**96.8**	**95.7**	**95.4**	**95.5**

6. Conclusion and Future Work

In this research, we proposed a novel framework, for the auto detection of student engagement in online learning using multi-modality. Unlike existing methods that depend primarily on visual cues and static classification, our approach integrates multiple modalities that include visual, audio, physiological, and interaction-based signals. We employed autoencoder-based non-linear embeddings that preserve complex feature dependencies. Engagement was modeled as a dynamic sequential process using temporal deep networks such as LSTMs, TCNs, and Transformers. An attention-based fusion strategy was implemented to adaptively weight modalities, enhancing robustness under diverse conditions. Through extensive empirical tests on four benchmarks datasets, DAiSEE, EmotiW, SEED-IV, and IIITB Online SE, it is shown that our framework outperforms existing state-of-the-art baseline models. Specifically, the proposed model achieves accuracy up to 96.8% and precision, recall, and F1-scores above 95% in all tests. The improvements were significant in the minority engagement classes which is seen as a longstanding problem/challenge in the field. These results confirm the effectiveness of combining multi-modal fusion, non-linear representation learning, and temporal modeling for reliable engagement detection. Future work will explore the inclusion of XAI methods to infer interpretable insights into engagement detection, allowing educators to comprehend which cues affect predictions and also, extending the framework for cross-cultural and cross-lingual datasets will promote fairness and generalization across different learning settings.

Author Contributions: Neella Swapna conceptualized the study, designed the multi-modal fusion framework, and led the development of the temporal deep network architecture for student engagement detection. She was primarily responsible for data preprocessing, feature extraction across modalities, experimental implementation, and performance evaluation. B. Ashwin Kumar contributed to the formulation of the research methodology, assisted in model validation and comparative analysis, and provided critical insights into result interpretation. Both authors collaboratively contributed to manuscript drafting, revision, and final approval of the submitted version.

Data availability: Data available upon request.

Conflict of Interest: There is no conflict of Interest.

Funding: The research received no external funding.

Similarity checked: Yes.

References

[1] S. Ikram, H. Ahmad, N. Mahmood, C. M. N. Faisal, Q. Abbas, I. Qureshi, and A. Hussain, "Recognition of student engagement state in a classroom environment using deep and efficient transfer learning algorithm," Applied Sciences, vol. 13, no. 15, p. 8637, 2023.

[2] O. M. Nezami, M. Dras, L. Hamey, D. Richards, S. Wan, and C. Paris, "Automatic recognition of student engagement using deep learning and facial expression," arXiv preprint arXiv:1808.02324, 2018.

[3] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School engagement: Potential of the concept, state of the evidence," Review of Educational Research, vol. 74, no. 1, pp. 59–109, 2004.

[4] E. L. Deci and R. M. Ryan, "The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior," Psychological Inquiry, vol. 11, no. 4, pp. 227–268, 2000.

[5] R. M. Carini, G. D. Kuh, and S. P. Klein, "Student engagement and student learning: Testing the linkages," Research in Higher Education, vol. 47, no. 1, pp. 1–32, 2006.

[6] S. Wang and Q. Li, "Student engagement in online learning: A review," Frontiers in Psychology, vol. 12, p. 7221, 2021.

[7] M. M. Santoni, T. Basaruddin, K. Junus, and O. Lawanto, "Automatic detection of students' engagement during online learning: A bagging ensemble deep learning approach," IEEE Access, vol. 12, pp. 96063–96073, 2024.

[8] EmotiW Engagement Prediction Challenge, available: https://sites.google.com/view/emotiw2020, accessed Jul. 2024.

[9] SEED-IV Dataset, BCMI Laboratory, Shanghai Jiao Tong University. [Online]. Available: http://bcmi.sjtu.edu.cn/~seed/seed-iv.html, accessed Jul. 2024.

[10] Z. Zhang, Z. Li, H. Liu, T. Cao, and S. Liu, "Data-driven online learning engagement detection via facial expression and mouse behavior recognition technology," Journal of Educational Computing Research, vol. 58, no. 1, pp. 63–86, 2020.

[11] J. Liao, Y. Liang, and J. Pan, "Deep facial spatiotemporal network for engagement prediction in online learning," Applied Intelligence, vol. 51, no. 10, pp. 6609–6621, 2021.

[12] Y. Chen, J. Zhou, Q. Gao, J. Gao, and W. Zhang, "MDNN: Predicting student engagement via gaze direction and facial expression in collaborative learning," Computational Modeling in Engineering & Sciences, vol. 136, no. 1, pp. 381–401, 2023.

[13] A. Abedi and S. S. Khan, "Improving state-of-the-art in detecting student engagement with ResNet and TCN hybrid network," arXiv preprint arXiv:2104.10122, 2021.

[14] O. M. Nezami, M. Dras, L. Hamey, D. Richards, S. Wan, and C. Paris, "Automatic recognition of student engagement using deep learning and facial expression," arXiv preprint arXiv:1808.02324, 2018.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 770–778, 2016.

[17] S. Ikram, H. Ahmad, N. Mahmood, C. M. N. Faisal, Q. Abbas, I. Qureshi, and A. Hussain, "Recognition of student engagement state in a classroom environment using deep and efficient transfer learning algorithm," Applied Sciences, vol. 13, no. 15, p. 8637, 2023.

[18] N. R. Paidja and F. A. Bachtiar, "Engagement emotion classification through facial landmark using convolutional neural network," in Proc. 2nd Int. Conf. Inf. Technol. Educ. (ICITE), pp. 234–239, 2022.

[19] M. M. Santoni, T. Basaruddin, K. Junus, and O. Lawanto, "Automatic detection of students' engagement during online learning: A bagging ensemble deep learning approach," IEEE Access, vol. 12, pp. 96063–96073, 2024.

[20] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2009 emotion challenge," in Proc. INTERSPEECH, pp. 312–315, 2009.

[21] EmotiW Engagement Prediction Challenge, available: https://sites.google.com/view/emotiw2020, accessed Jul. 2024.

[22] SEED-IV Dataset, BCMI Laboratory, Shanghai Jiao Tong University. [Online]. Available: http://bcmi.sjtu.edu.cn/~seed/seed-iv.html, accessed Jul. 2024.

[23] A. Abedi, C. Thomas, D. B. Jayagopi, and S. S. Khan, "Bag of states: A non-sequential approach to video-based engagement measurement," arXiv preprint arXiv:2301.06730, 2023.

[24] I. Dubovi, "Cognitive and emotional engagement while learning with VR: The perspective of multimodal methodology," Computers & Education, vol. 183, p. 104495, 2022.

[25] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, 2019.

[26] O. M. Nezami, M. Dras, L. Hamey, and D. Richards, "Deep learning for engagement detection in online learning: A two-stage transfer learning approach," in Proc. Int. Conf. Adv. Learn. Technol. (ICALT), pp. 406–410, 2019.

[27] T. Selim, I. Elkabani, and M. A. Abdou, "Students engagement level detection in online e-learning using hybrid EfficientNetB7 together with TCN, LSTM, and Bi-LSTM," IEEE Access, vol. 10, pp. 99573–99583, 2022.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] P. Buono, B. De Carolis, F. D'Errico, N. Macchiarulo, and G. Palestra, "Assessing student engagement from facial behavior in online learning," Multimedia Tools and Applications, vol. 82, no. 9, pp. 12859–12877, 2023.

[30] A. Abedi and S. S. Khan, "Improving state-of-the-art in detecting student engagement with ResNet and TCN hybrid network," arXiv preprint arXiv:2104.10122, 2021.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 5998–6008, 2017.

[32] https://iith.ac.in/~daisee-dataset/

[33] https://sites.google.com/site/emotiw2014/

[34] https://bcmi.sjtu.edu.cn/home/seed/

[35] https://iiitb.ac.in/research/data-sets