Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Multimodal Fusion Framework for Student Engagement Detection Using Facial, Behavioral, and Textual Features

Fatima Khatib
COEP Technological University
Soma Ghosh
gsn.comp@coeptech.ac.in

COEP Technological University
Suraj Sawant
COEP Technological University
Manjari Gupta
SplashBI

Research Article

Keywords: Student engagement, multimodal fusion, action units, XGBoost, deep learning

Posted Date: September 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10801110/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Multimodal Fusion Framework for Student Engagement Detection Using Facial, Behavioral, and Textual Features

Fatima Khatib[1], Soma Ghosh[1*], Suraj Sawant[1], Manjari Gupta[2]

[1*]Department of Computer Science and Engineering,
COEP Technological University Pune, Maharashtra, 411005, India.
[2]SplashBI, Toronto, Ontario, Canada.

*Corresponding author(s). E-mail(s): gsn.comp@coeptech.ac.in;
Contributing authors: khatibfv24.comp@coeptech.ac.in;
sts.comp@coeptech.ac.in; manjarigupta0909@gmail.com;

Abstract

Student engagement in classrooms has been shown to enhance motivation and improve academic performance; however, maintaining student engagement during instruction remains a significant challenge for educators. This study presents a multimodal framework for student engagement detection by integrating facial, behavioral, and textual cues derived from images, videos, and text data. The proposed framework investigates the contribution of individual modalities and evaluates both feature-level (early) and decision-level (late) fusion strategies for engagement prediction. A key contribution of this work is the analysis of modality-specific influences on engagement detection, providing practical insights for educational applications. Publicly available datasets comprising image, video, and textual data were employed to classify student engagement into three categories: engaged, partially engaged, and disengaged. Experimental results indicate that XGBoost achieved the highest accuracy of 0.867 using handcrafted behavioral features, highlighting the predictive significance of observable actions and eye-movement patterns. Furthermore, the late fusion model integrating all three modalities achieved an accuracy of 0.864, outperforming individual modality-based models. The findings demonstrate the effectiveness of multimodal fusion for student engagement detection and highlight its potential for developing intelligent classroom monitoring systems.

Keywords: Student engagement, multimodal fusion, action units, XGBoost, deep learning

1 Introduction

Classroom involvement is essential for students to be successful academically and to achieve their full potential in the learning process. Student involvement indicates the degree to which students pay attention, have motivation, and use cognitive processing when participating in instructional events. When students are actively engaged in classrooms, they experience curiosity and take part in self-regulated learning activities [1]. Students who are not engaged find it difficult to maintain attention and achieve successful academic outcomes. Therefore, classroom involvement is not simply one observed behavior; it is a collection of behaviors that reflect cognitive, emotional, and behavioral states. The rise of e-learning environments and blended classroom models provides a new form of pstudent participation but makes it difficult for teachers to monitor it [2]. As a result, instructors rely on quizzes or assignments to assess student participation. These methods are subjective and provide limited information about the quality of engagement [2]. The use of artificial intelligence (AI) in education helps analyze engagement levels, identify at-risk students, and support efficient teaching practices [3]. Visual and behavioral signals such as facial expressions, gaze, and posture are most commonly used in such systems to estimate students' attention. Earlier work focused on the use of single modalities; the results of these indicated that these are not reliable indicators of engagement on their own [5]. Action Units (AUs), or facial muscle movements, are also used to capture behavioral and emotional aspects of student engagement [4]. Multimodal approaches can address these limitations by using input signals from multiple data sources such as combining facial expressions, audio transcripts, and physiological signals to capture more nuances in engagement.

In this research, we present a unified framework that investigates the usage of textual data in conjunction with image and video data. Cognitive and emotional engagement cues that are not captured by image or video data alone are captured by textual data. The DAiSEE [7] dataset and the WACV student engagement dataset [6] are used for image and video clips. MediaPipe [8] and OpenFace [9] are used to extract features from these datasets. Semantic embeddings are used to extract textual features from the IntrEx dataset [10], which contains student-teacher interaction data. In contrast to previous research, we examine early and late fusion procedures within both frameworks and offer a thorough comparison of machine learning (ML) and deep learning (DL) models.

2 Related Works

A number of issues, such as dataset quality, multimodal fusion, and real-time deployment, remain unaddressed despite notable advancements in e-learning and collaborative learning settings. In order to assess engagement, studies by [11–13] mostly depend on visual characteristics like AU, eye gaze, facial landmarks, and head pose. Specifically, students' cognitive states may not be well captured by [11], which relies solely on gaze information to infer engagement. Additionally, these studies only give a partial picture of student participation due to the exclusion of textual data. Multimodal strategies that combine video and audio transcripts have been investigated, however they don't always result in better performance. For example, [14] found that

their combined speech-video model performed worse than unimodal models, indicating flaws in the fusion strategy used. Furthermore, the use of synthetic or small-scale datasets in [13, 15] raises questions about how well these methods translate to real-world classroom environments, where issues like occlusions, background noise, and tracking multiple students at once are common [16].

Chat messages, Learning Management System (LMS) logs, and other student interaction data are used by the authors in [17–19]. Because they take data from specific platforms, this reduces the model's capacity to generalize to other datasets. Subjectivity is introduced by students' self-reporting of their involvement in these databases. Since [18] only assesses engagement after students finish a task, it is difficult to monitor students' level of participation during an activity. In a similar vein, [19] notes that a model's performance is determined on the quality of annotations added to training data. The data's ambiguous labeling causes some kids to be incorrectly classed. Additionally, the authors contend that there is a great deal of promise for using unstructured text data (from social media, etc.), but that this data has to be better connected with other modalities. Performance can be significantly impacted by fusion. Feature imbalances and feature noise, where features fluctuate by scale and/or degree of reliance depending on each modality used, are common problems in early fusion. The link between modalities is not always captured by late fusion techniques. Poor time-synchronization and poor alignment between modalities lead to poor performance, as demonstrated by [14] and [19]. Attention-based and transformer-based fusion models, which are detailed in [18, 20], are highly computationally costly. These don't work well in practical applications since they need a lot of training data, which aren't always easily accessible.

Extracting visual attributes is the main focus of most studies. Textual and interaction-based data have either not been thoroughly examined or have been poorly integrated by researchers. The majority of current approaches provide less-than-ideal outcomes because they do not adequately understand intermodal interactions. The majority of datasets restrict generalizability because they are either too small, oriented toward a specific demographic, or subjectively annotated. Few research have used the same set of input features to compare ML and DL techniques. Unlike earlier research, this work use a variety of modalities to identify both the underlying cognitive involvement and the observable behaviors. Additionally, this study shows the advantages and disadvantages of early versus late fusion. The same feature sets are used to test both ML and DL models. In addition to addressing the problems associated with multi-modality integration, this study contributes to the improvement of engagement detection systems appropriate for educational settings.

3 Proposed Methodology

The overall architecture of the proposed framework is illustrated in Figure 1. The process comprises dataset preparation, feature extraction, feature representation, feature fusion, and classification. Each of these stages is described in detail in the subsequent sections.

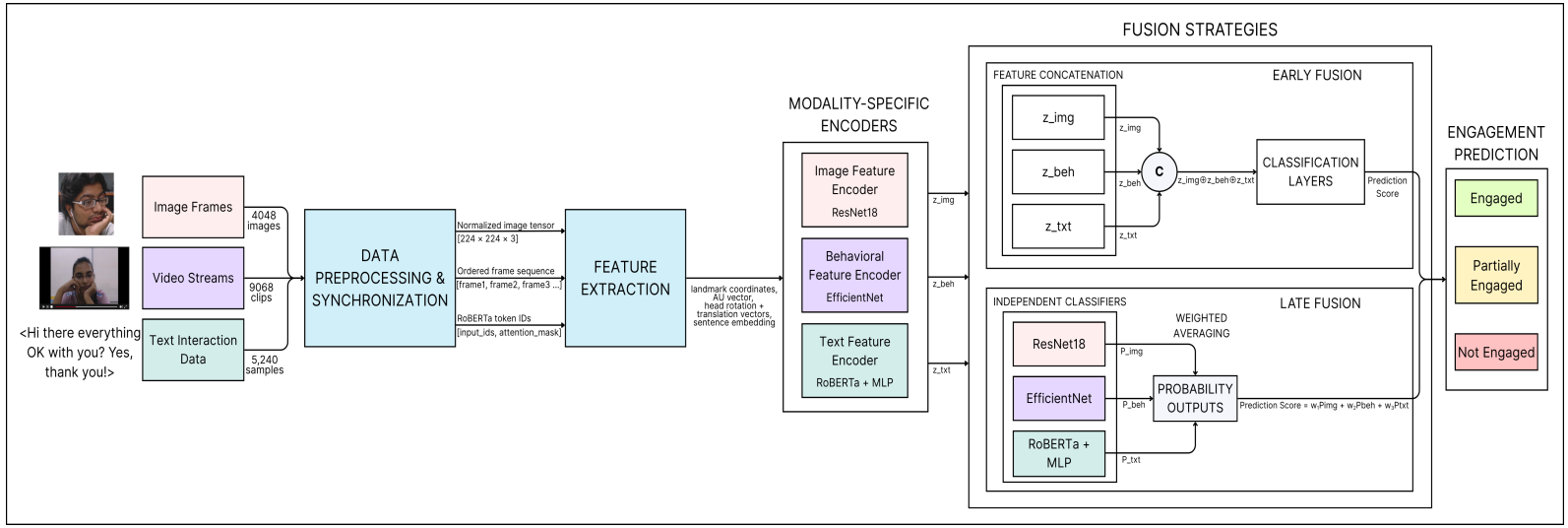


Fig. 1 Overall architecture of the proposed framework.

Table 1 Interestingness Score Mapping to Engagement Labels

INT Range	Engagement Label	Class	No. of samples
0-1	Disengaged	0	3366
2	Partially Engaged	1	1962
3-4	Engaged	2	473

Table 2 Summary of Extracted Features

Feature Type	No. of Features	Dataset
Eye Gaze	6 direction vectors	DAiSEE
Head Pose	6 rotation & location vectors	DAiSEE
Facial Landmarks	468 key point coordinates	WACV
Action Units	17 muscle movements	DAiSEE
OpenFace Features	712 features (concat.)	WACV, DAiSEE
Text Embeddings	5801 samples (768-dim each)	IntrEx

3.1 Dataset

The WACV dataset consists of 4,424 annotated images categorized into three levels of engagement: disengaged (412 samples), partially engaged (2,247 samples), and fully engaged (1,765 samples). This class distribution reflects realistic classroom settings, where completely disengaged students are observed less frequently than partially or fully engaged students.

The DAiSEE dataset comprises 9,068 annotated video clips, each 10 seconds long, capturing the temporal variations in student engagement. To further incorporate the cognitive dimension of engagement, the IntrEx dataset is used. This dataset contains 5,801 conversational segments represented as sequential teacher–student dialogues. Each dialogue segment is independently annotated by three raters for the level of interest (INT) on a scale ranging from 0 to 4.

In Equation (1), $p0_{int}$, $p1_{int}$, and $p2_{int}$ denote the interest ratings assigned by the three annotators. The final interest score, denoted by INT_{final}, is computed as the average of these ratings and serves as the ground-truth measure for classifying engagement. Subsequently, the INT_{final} scores are discretized into three engagement classes, as summarized in Table 1.

$$INT_{final} = \frac{p0_{int} + p1_{int} + p2_{int}}{3} \tag{1}$$

3.2 Data Preprocessing

To mitigate unwanted variations in the image data, several preprocessing steps are applied. First, all images are resized to a uniform resolution of $224 \times 224 \times 3$ and normalized using the mean and standard deviation values derived from ImageNet. Data augmentation techniques, including random horizontal flipping and rotation, are employed to improve model generalization and robustness. For the DAiSEE

video dataset, frames are uniformly sampled to capture temporal variations in student behavior. Furthermore, special tokens (e.g., <STUDENT>) are preserved in the textual data to retain contextual information associated with teacher–student interactions, thereby facilitating more effective engagement detection.

3.3 Feature Extraction and Representation

The Facial Action Coding System (FACS) defines Action Units (AUs) as specific facial muscle activations or movements. Complex facial expressions are represented by the simultaneous activation of multiple AUs. In this study, OpenFace is employed to extract facial behavioral cues, including eye gaze and head pose, by estimating the direction of a student's gaze from facial landmarks and determining the orientation of the head with respect to the camera.

For the textual modality, contextual embeddings are generated using the RoBERTa-base model. Specifically, RoBERTa encodes each dialogue sample into a dense sentence-level representation by extracting token-based contextual features. RoBERTa is selected due to its ability to effectively capture semantic information, sentiment, intent, and linguistic cues indicative of student participation. These contextual representations facilitate the modeling of temporal variations in language associated with different levels of engagement.

3.4 Machine Learning Models

Five machine learning algorithms were evaluated: Support Vector Machines (SVM), Random Forest (RF), Gradient Boosting, Decision Trees (DT), and eXtreme Gradient Boosting (XGBoost). These algorithms process various combinations of input features, enabling systematic comparison of the contributions of individual modalities. Four feature sets were examined: (S1) facial landmark-based features; (S2) head pose and eye gaze-based features; (S3) Action Unit (AU)-based features; and (S4) a combined feature set incorporating S1–S3. Given the limited number of training samples, relatively simple deep learning architectures were employed rather than complex models. Lightweight feature extraction methods and moderate-sized architectures reduce computational requirements for both training and inference. Average inference times remain near real-time, facilitating straightforward integration into online classroom environments or learning analytics platforms.

3.5 Deep Learning Models

The image-based model operates on facial landmark features extracted from the WACV dataset and employs a ResNet18 architecture pre-trained on the ImageNet dataset, followed by a task-specific classification layer. For the behavioral modality, Action Units (AUs), eye gaze, and head pose features extracted from the DAiSEE dataset are processed using an EfficientNet-based neural network.

For the textual modality, contextual embeddings are provided as inputs to a Multi-Layer Perceptron (MLP) classifier. The MLP architecture is selected because of its

Table 3 Feature Weights for Late Fusion

Feature	Weight Value
Facial Landmark (w_1)	0.3
Activation Units (w_2)	0.2
Eye Gaze & Head Pose (w_3)	0.2
Dialogue Embeddings (w_4)	0.3

relatively simple structure compared with deeper neural architectures, thereby reducing model complexity and minimizing the risk of overfitting while maintaining effective classification performance.

3.6 Fusion

Early Fusion: A unified feature representation is obtained by concatenating the feature vectors extracted from the three modalities into a single embedding vector. The resulting concatenated representation is subsequently processed through fully connected layers with dropout regularization. In Equation (2), X_{img}, X_{beh}, and X_{txt} denote the encoded image, behavioral, and textual features, respectively, which are combined to form the joint representation. This fusion strategy facilitates the learning of cross-modal interactions; however, its performance may be adversely affected by noisy or unreliable features of individual modalities.

$$X = [X_{img}; X_{beh}; X_{txt}] \tag{2}$$

Late Fusion: In the late fusion strategy, each modality-specific model is trained independently to produce a class probability distribution. Specifically, P_{img}, P_{AU}, P_{EH}, and P_{txt} denote the probability vectors generated by the facial landmark, behavioral, and textual models, respectively. The final prediction, P_{final}, is obtained by computing a weighted average of these modality-specific probability distributions, as defined in Equation (3). The weights w_1, w_2, w_3, and w_4 correspond to the contributions of facial landmarks, Action Units (AUs), gaze and head pose, and text embeddings, respectively. This fusion approach enables each modality to contribute to the final decision based on its relative importance.

$$P_{final} = w_1 P_{img} + w_2 P_{AU} + w_3 P_{EH} + w_4 P_{txt} \tag{3}$$

Modality-specific weights are optimized via a grid search in the validation set, subject to the constraints specified in Equation (4). Once the optimal weights are identified, they are used to generate the final predictions on the test set. The weight combination that yields the highest weighted F1-score on the validation set is selected as the optimal configuration and reported in Table 3.

$$w_1 + w_2 + w_3 + w_4 = 1 \tag{4}$$

Table 4 Hyperparameters Across Modalities

Model	LR	Batch Size	Dropout
ResNet (IF)	$2.54e^{-4}$	64	0.5
EfficientNet (BF)	$1.32e^{-4}$	32	0.5
MLP (TF)	$1e^{-3}$	32	0.3
IF+BF	$1.58e^{-4}$	32	0.3-0.5
BF+TF	$1e^{-3}$	32	0.3-0.5
IF+TF	$2.54e^{-4}$	32	0.3-0.5
IF+BF+TF (Early Fusion)	$1.32e^{-4}$	32/64	0.2-0.5
IF+BF+TF (Late Fusion)	$1e^{-4}$	32/64	0.2-0.5

4 Results

4.1 Experimental Setup

To accommodate the heterogeneous dimensionalities of multimodal features, the fusion models are configured with flexible parameter settings. All models are trained for a maximum of 50 epochs. To prevent overfitting, early stopping with a patience of 30 epochs, based on validation loss, is employed for the fusion models. The learning rate is adaptively selected according to the complexity of the input features and the architecture of the deep learning model. The complete set of hyperparameter configurations used for training is summarized in Table 4.

The datasets are divided into training, validation, and test subsets in a ratio of 70:15:15. To address class imbalance, class-weighted loss functions are incorporated during model training. The class weight, denoted by w_i, is computed using Equation (5), where N_i represents the number of samples belonging to the ith engagement class.

$$w_i = \frac{1}{N_i} \tag{5}$$

4.2 Model Performance

Among the evaluated machine learning models, XGBoost achieved the best performance when trained on the combined feature set, attaining an accuracy of 0.867 and an F1-score of 0.871. This result indicates that integrating geometric facial features, gaze patterns, head pose, and Action Units (AUs) yields a more comprehensive representation of student engagement than any single feature modality. The performance comparison of the machine learning models is presented in Table 5.

In contrast, the Support Vector Machine (SVM) exhibited a notable decline in performance, suggesting that its decision boundaries are insufficient to effectively capture the complexity and nonlinearity inherent in the fused feature space. Among the unimodal feature sets, AUs demonstrated the highest discriminative capability, achieving an accuracy of 0.864. This finding highlights the significance of facial muscle activations in representing the cognitive and affective states associated with student engagement.

Facial landmark- and gaze-based features exhibit comparatively lower performance than other feature modalities. This reduced performance can be attributed to their sensitivity to various sources of noise, including illumination variations, occlusions,

Table 5 Evaluation of all ML Models for 4 Feature Sets

Features	Model	Accuracy	Precision.	Recall	F1 Score
Facial landmarks	SVM	0.612	0.624	0.608	0.61
	RF	0.842	0.848	0.839	0.841
	DT	0.821	0.827	0.818	0.82
	GBM	0.851	0.856	0.847	0.849
	XGBoost	**0.859**	**0.864**	**0.857**	**0.862**
Gaze+Pose	SVM	0.598	0.603	0.594	0.596
	RF	0.826	0.831	0.823	0.825
	DT	0.814	0.819	0.81	0.812
	GBM	**0.856**	**0.868**	**0.863**	**0.865**
	XGBoost	0.844	0.85	0.839	0.842
AU	SVM	0.621	0.629	0.616	0.618
	RF	0.846	0.852	0.843	0.845
	DT	0.829	0.835	0.826	0.828
	GBM	0.852	0.86	0.85	0.851
	XGBoost	**0.864**	**0.878**	**0.864**	**0.866**
Combined	SVM	0.635	0.642	0.631	0.633
	RF	0.853	0.86	0.849	0.851
	DT	0.838	0.843	0.834	0.836
	GBM	0.861	0.866	0.857	0.859
	XGBoost	**0.867**	**0.873**	**0.866**	**0.871**

and head movements. To further investigate the effectiveness of each modality, the deep learning (DL) models are evaluated independently, with the results summarized in Table 6. Among the unimodal DL models, the behavioral model achieves the highest accuracy of 0.823, followed by the text-based model (0.817) and the image-based model (0.809). These findings suggest that structured behavioral features, such as Action Units (AUs), eye gaze, and head pose, are more informative for engagement detection than facial landmark-based representations. Unlike raw image data, behavioral features are less susceptible to environmental noise and are better able to capture meaningful engagement-related patterns.

The text-based model demonstrates moderate performance. Although textual interactions provide valuable insights into cognitive engagement, the presence of ambiguous, short, or low-information utterances may limit their discriminative capability when considered in isolation. Consequently, textual features alone are insufficient to comprehensively characterize student engagement.

All multimodal DL combinations outperform their corresponding unimodal counterparts. As shown in Table 7, the combination of behavioral and textual modalities achieves the highest performance, with an accuracy of 0.845. This result indicates that behavioral features capture observable manifestations of attention and affect, whereas textual features reflect cognitive aspects of engagement. The integration of these complementary modalities, therefore, enhances engagement recognition performance. The inclusion of image-based features yields only marginal improvements, likely due to the substantial variability and noise inherent in visual data. Table 7 further presents the performance of the proposed fusion strategies utilizing all three modalities.

Early fusion combines features from multiple modalities at the representation level prior to classification. However, multimodal systems often comprise heterogeneous

Table 6 Performance of individual deep learning models

Features	Model	Accuracy	Precision	Recall	F1 Score
IF	ResNet	0.809	0.813	0.807	0.804
BF	EfficientNet	0.823	0.826	0.823	0.829
TF	MLP	0.817	0.827	0.816	0.812

Table 7 Performance of bimodal and trimodal models using early and late fusion

Model	Fusion	Accuracy	Precision	Recall	F1 score
IF + BF	Early	0.834	0.834	0.838	0.832
IF + TF	Early	0.83	0.831	0.827	0.824
BF + TF	Early	0.845	0.857	0.846	0.842
IF + BF + TF	Early	0.856	0.864	0.853	0.858
IF + BF + TF	Late	0.864	0.873	0.867	0.861

data sources with distinct probability distributions, which can limit the effectiveness of feature-level integration and hinder the learning of a unified representation. In contrast, late fusion employs independent models for each modality to learn modality-specific representations, then combines their output probabilities to produce the final prediction.

Several important observations can be drawn from the experimental results. First, behavioral features are the most reliable indicators of student engagement, as they provide structured, interpretable representations of engagement-related cues. Second, textual data effectively capture cognitive aspects of engagement, particularly in situations where visual indicators of attention are weak or ambiguous. In contrast, facial landmark-based approaches exhibit lower robustness, primarily due to their sensitivity to external factors such as illumination variations, occlusions, and changes in head orientation relative to the camera.

The results obtained from multimodal fusion further confirm that student engagement is inherently multidimensional, encompassing behavioral, cognitive, and affective components. Among the evaluated fusion strategies, early fusion consistently outperforms late fusion, particularly under low-quality visual input conditions. Nevertheless, despite the overall strong predictive performance, the proposed models remain susceptible to classification errors. False-positive predictions are frequently observed when students appear visually attentive but are cognitively disengaged, leading models to overestimate engagement levels. Conversely, false negatives occur when students exhibit limited visual attention cues, for example, while taking notes or reflecting deeply on the instructional content, causing the models to incorrectly classify them as disengaged. In addition, ambiguities in textual responses contribute to misclassification, especially when students provide short or generic responses, such as "yes" or "ok", which convey limited information about their actual state of engagement.

4.3 Statistical Validation

Model performance is evaluated using multiple metrics, including accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). In addition, the discriminative capability of the models trained on the combined feature set is assessed across different decision thresholds using Receiver Operating Characteristic (ROC) curves, as illustrated in Figure 2. Among the evaluated machine learning models, XGBoost and Gradient Boosting exhibit superior discriminative performance, achieving AUC values of 0.99 and 0.98, respectively.

Given that the late fusion strategy achieved slightly higher performance (accuracy of 0.864) than early fusion, its training dynamics were further analyzed over 50 epochs. The learning curves indicate that the model effectively learns from the integrated multimodal representations. As shown in Figure 3, the training accuracy increases steadily, reaching approximately 90%, while the validation accuracy stabilizes in the range of 86–87%. The corresponding loss curves, presented in Figure 4, exhibit an exponential decline, with the minimum training loss attained around epoch 30. To mitigate overfitting and prevent the model from adapting excessively to noise present in the training data, early stopping is employed at epoch 30. Optimal model performance is achieved when the validation loss reaches its minimum, and the validation accuracy converges to a stable plateau.

To evaluate generalization capability, 5-fold cross validation is performed on the unimodal and multimodal deep learning models. In each fold, the dataset is divided into training and testing subsets such that four folds are used for training and one fold is used for testing. The process is repeated five times, and the final performance is computed as the average across all folds. Accuracy, precision, recall, F1-score, and

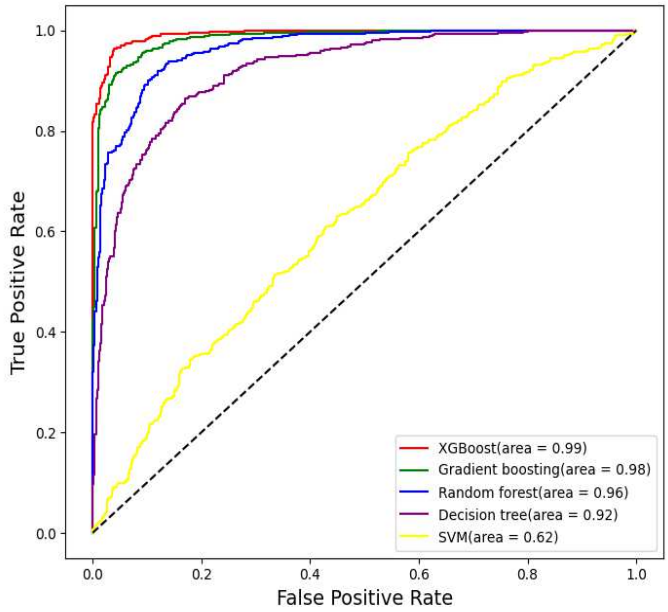


Fig. 2 Receiver Operating Curve (ROC) for Combined Features

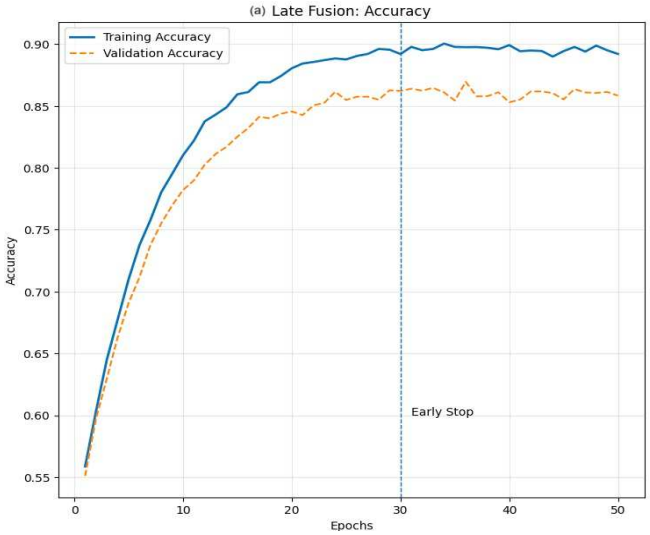


Fig. 3 Training and Validation Accuracy Curve

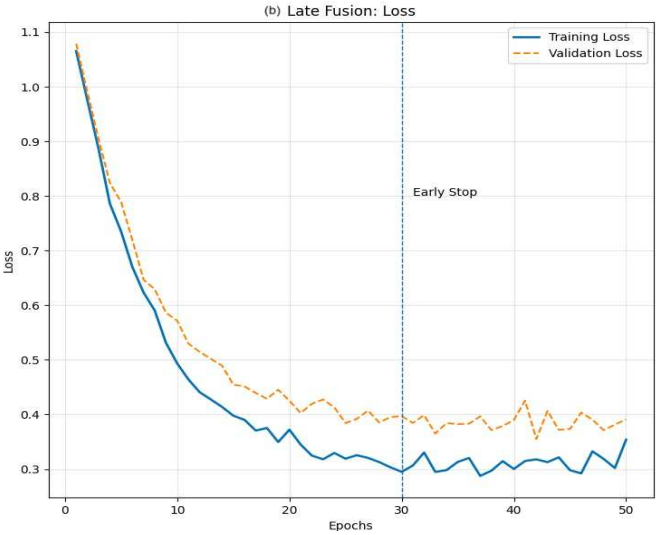


Fig. 4 Training and Validation Loss for Late Fusion

standard deviation are used as evaluation metrics in order to measure both predictive capability and consistency across different data partitions.

The behavioral model achieved the strongest unimodal performance with an average accuracy of $82.3\% \pm 0.6$, followed by the text model and image model. Early fusion achieved an average accuracy of $85.6\% \pm 0.4$, while the proposed late fusion framework achieved the highest average accuracy of $86.4\% \pm 0.3$ and an F1-score of $86.1\% \pm 0.4$. The lower standard deviation observed in the late fusion framework in

shown in Table 8. Stable learning behavior and improved robustness is observed across varying train-test splits.

Paired t-tests are conducted using the fold-wise accuracy scores, displayed in Table 9. The comparison between early fusion and late fusion produced a p-value of 0.031, indicating statistically significant improvement ($p < 0.05$). Comparison between the behavioral model and late fusion produced a p-value of 0.004, demonstrating highly significant improvement ($p < 0.01$). Statistical evaluation confirms that fusion maintains consistent and reliable predictive behavior across multiple validation folds.

Table 8 Statistical validation using 5-fold cross validation

Model	Accuracy (%)	F1 Score (%)	Standard Deviation
Behavioral Model	82.3 ± 0.6	82.7 ± 0.5	0.006
Early Fusion	85.6 ± 0.4	85.8 ± 0.3	0.004
Late Fusion	86.4 ± 0.3	86.1 ± 0.4	0.003

Table 9 Paired t-test results between models

Comparison	P-value	Statistical Significance
Early Fusion vs Late Fusion	0.031	Significant ($p < 0.05$)
Behavioral vs Late Fusion	0.004	Highly Significant ($p < 0.01$)

5 Comparative Analysis

Most existing studies focus on a single input modality, while only a limited number have explored the integration of all three modalities. Table 10 presents a detailed comparison between the proposed approach and existing methods. The authors in [11] reported an accuracy of 94.8% using facial expression and gaze-based features. Emotion detection was incorporated to identify disengaged students. Nevertheless, the study did not consider linguistic cues or other communication modalities, limiting its multimodal representation. The mothod proposed in [12] used both ML and DL models for detecting student engagement. Their best results were achieved using ML models that analyzed facial and behavioral characteristics, whereas DL performed much lower (only 47.2%). The framework presented in [23] achieved an accuracy exceeding 95% using a combination of facial and gait-based biometric features. Nevertheless, the reliance on gait-based biometrics poses challenges for deployment in classroom settings, thereby limiting the model's generalizability. In addition, the framework is primarily designed to analyze emotional states and does not offer a comprehensive multimodal assessment of student engagement. [21] demonstrated that integrating multiple modalities yielded higher accuracy than approaches based on a single modality. However, the generalizability of the proposed framework across diverse classroom environments and learning scenarios may be limited due to the relatively small dataset

comprising only 1,145 samples. Similarly, [19] developed a multimodal framework that combined video, text, and log data. Although promising results were reported, the study involved only 27 participants, limiting the model's ability to effectively distinguish between highly and less engaged students.

Furthermore, [18] demonstrated the effectiveness of integrating chat and facial modalities, achieving improvements over unimodal baselines. Nevertheless, the reliance on post-activity satisfaction ratings as ground truth constrains the applicability of the framework for real-time engagement assessment and analysis. In addition, the limited dataset of only 164 samples further restricts the generalizability and robustness of the proposed system

Advanced approaches, including multimodal transformers and cross-modal attention mechanisms, generally demand extensive training datasets, significant computational resources, and careful model tuning. Furthermore, attention-based fusion techniques have been employed to integrate physiological sensor data with visual modalities. Despite achieving high prediction accuracies, the practical deployment of such methods in resource-constrained classroom environments is challenging. The proposed framework offers competitive predictive performance relative to existing multimodal systems while maintaining interpretability

The experimental results demonstrate that multimodal fusion consistently outperforms individual modalities. Specifically, the weighted late fusion approach achieved an average accuracy of 0.864 and an average F1-score of 0.861, yielding performance that is competitive with, or superior to, existing multimodal systems while maintaining a balance between predictive capability and interpretability. This improvement can be attributed to the complementary nature of the employed modalities. Behavioral features capture attentional and engagement-related cues more effectively than visual features or text embeddings alone, whereas text embeddings provide contextual and semantic information regarding student interactions that cannot be inferred solely from behavioral data. By integrating these modalities within a unified framework, a richer and more discriminative feature space can be leveraged than would be possible by relying exclusively on either behavioral or textual features alone. Furthermore, weighted late fusion preserves the strengths of individual modalities while mitigating the influence of noise introduced by weaker modalities.

6 Conclusion and Future Work

This study presents a multimodal framework for student engagement detection based on facial, behavioral, and textual cues. The experimental results demonstrate that integrating visual, behavioral, and textual information improves engagement detection performance compared with unimodal approaches. Both feature-level (early) fusion and decision-level (late) fusion strategies were investigated, with fusion-based approaches consistently outperforming individual modality models.

Among the evaluated classifiers, XGBoost achieved the highest overall accuracy of 0.867 using handcrafted behavioral features, highlighting the continued predictive importance of observable actions and eye-movement patterns. Comparative analysis

Table 10 Comparison of existing student engagement detection methods with the proposed system

Ref No.	Datasets	Modality	Features	Models	Accuracy	Limitation
[17]	Posts, comments, questionnaire responses from 51 students	Text	Number sequence, TF-IDF, Word2Vec	RF, SVM, FCNN, LSTM	0.886	Dataset is limited to one discipline; binary classification of engagement
[21]	1,145 instances of gameplay data from 31 students, FER-2013 dataset	Image and Text	Facial action units, emotion probabilities, gameplay logs	kNN, DT, RF, LR with decision-level fusion	0.836	No student self-report validation; tied to one puzzle game; limited fusion strategies
[14]	RAVDESS dataset	Video and audio	Facial emotion cues, acoustic speech features	CNN models (speech, video, combined multimodal)	Video: 0.947, speech: 0.928, combined: 0.693	Combined model performed worse than unimodal models
[22]	Data from 32 students using IoT sensors and therapy sessions	Sensor data, text	Educational data, environmental data, physiological data, motion sensor data	kNN, DT, RF, NB, MLP, SVM, XGBoost, DNN	0.977	Binary prediction limits insights; data collected only in one-to-one therapy setting
[11]	Student Engagement Dataset, AffectNet	Image	Facial action units,15 gaze vectors, emotion indicators	CNN, SVM, DNN	0.948	Self-reported labels; Dataset labels focus mainly on gaze; emotion detection accuracy is relatively low

Table 10 Comparison of existing student engagement detection methods with the proposed system (PART II - Continued)

Ref No.	Datasets	Modality	Features	Models	Accuracy	Limitation
[20]	DECEP	Facial features, Sensor data	EEG, eye tracking, skin conductance, blood volume changes, facial features	Transformer, DNN, Bi-LSTM	Mean Absolute Error: 0.3544, Mean Square Error: 0.1651	Small participant pool; self-reported labels
[16]	DEAP	Sensor data	EEG frequency sub-bands	1D-CNN, LIME	0.999	class imbalance; binary emotion detection
[13]	MSR, NPUAction, classroom dataset	Video	Facial expressions, abnormal behaviors, head posture	VGG16, CNN	0.945	Modality imbalance; subjective labeling
[23]	Biometric and sentiment datasets	Image, gait	Face and gait recognition, facial expressions	PCA, attention-based models	0.851	Dataset not public; occlusions and crowded scenes
[18]	activity instances	Text and video	Chat messages, facial AUs, head pose, eye gaze	LSTM, cross-attention model	0.72	Satisfaction measured post-activity; individual responses only; averaging facial features may skew results

Table 10 Comparison of existing student engagement detection methods with the proposed system (PART III - Continued)

Ref No.	Datasets	Modality	Features	Models	Accuracy	Limitation
[15]	Five public facial emotion datasets	Image and Video	Facial expressions, facial landmarks, lip movement	Haar cascade, CNN, Dlib, multimodal feature aggregation	Engagement detection: 0.935, Emotion detection: 0.99	No group analysis; limited dataset for engagement evaluation
[12]	WACV and DAiSEE	Image, Video	AUs, gaze, head pose, facial landmarks	SVM, RF, DT, XGBoost, ResNet, EfficientNet	ML: 0.829, DL: 0.472	Only visual modalities; dataset biaseds
[19]	Dataset from 27 students	Video and Text	Facial expressions, chat messages, behavioral logs	TimeCNN, FCN, Encoder, MCNN	0.95	Annotation accuracy affects performance; difficulty detecting high engagement.
Prop methodology	WACV, DAiSEE, Student text interaction data	Image, Behavioral, Text	Facial landmarks, AUs, eye gaze, head pose, SBERT sentence embeddings	SVM, RF, DT, GBM, ResNet18 (image), EfficientNet (behavioral), MLP (text) with early and late fusion	**Image: 0.809, behavioral: 0.823, text: 0.817, early fusion: 0.856, late fusion: 0.864**	no temporal analysis.

further revealed that late fusion outperformed early fusion, suggesting that decision-level integration is more effective in handling heterogeneous feature representations. Although behavioral features emerged as the strongest individual modality, combining behavioral and textual modalities yielded the greatest improvement in predictive performance. This finding indicates that while behavioral cues reflect students' attentional states, textual information provides complementary insights into their cognitive processes. Consequently, integrating facial, behavioral, and textual modalities provides a more comprehensive assessment of student engagement.

Despite these promising results, the temporal evolution of engagement during video-based learning was not considered in the present study. As a result, model performance remains highly dependent on the quality and reliability of annotated training data. In particular, engagement studies that rely on textual data often suffer from unreliable ground-truth labels, underscoring the need for more robust and novel annotation strategies.

Future research will focus on incorporating explainable artificial intelligence (XAI) techniques to improve model transparency by identifying the factors contributing to engagement predictions. Additionally, developing flexible mechanisms for handling missing modalities will be essential for real-world deployment. Finally, temporal modeling approaches, such as Long Short-Term Memory (LSTM) networks and transformer-based architectures, will be explored to capture the dynamic nature of student engagement over time.

Declarations

Competing Interests

The authors declare no conflict of interest.

Funding Information

The authors did not receive support from any organization for the submitted work.

Clinical Trial Number

Not Applicable.

Data Availability

The datasets analysed during the current study are available in the following links:-

- https://people.iith.ac.in/vineethnb/resources/daisee/index.html
- https://github.com/e-drishti/wacv2016
- https://www.cl.cam.ac.uk/research/rainbow/projects/openface/
- https://huggingface.co/collections/XingweiT/intrex-68a8f2c97688157066860ae2.

Ethics Declaration

Not Applicable.

Consent to Participate

Not applicable.

Consent to Publish

All authors have read and approved the final version of this manuscript and consent to its publication.

References

[1] Zeng, J., Yan, H., & Zhang, H. (2025). How Classroom Curiosity Affects College Students' Creativity? Education Sciences, 15(9), 1101. https://doi.org/10.3390/educsci15091101

[2] Simon, P. D., Jiang, J., & Fryer, L. K. (2025). Assessment of class participation in online and offline learning environments: a qualitative investigation of teachers' best practices and challenges. Higher Education Research & Development, 44(5), 1191–1208. https://doi.org/10.1080/07294360.2025.2462024

[3] Chen, Xieling, Di Zou, Haoran Xie, Gary Cheng, and Caixia Liu. "Two decades of artificial intelligence in education." Educational Technology & Society 25, no. 1 (2022): 28-47.

[4] Sharma, Prabin, Shubham Joshi, Subash Gautam, Sneha Maharjan, Salik Ram Khanal, Manuel Cabral Reis, João Barroso, and Vítor Manuel de Jesus Filipe. "Student engagement detection using emotion analysis, eye tracking and head movement with machine learning." In International conference on technology and innovation in learning, teaching and education, pp. 52-68. Cham: Springer Nature Switzerland, 2022.

[5] Qi, Yongfeng, Liqiang Zhuang, Huili Chen, Xiang Han, and Anye Liang. "Evaluation of students' learning engagement in online classes based on multimodal vision perspective." Electronics 13, no. 1 (2023): 149.

[6] Kamath, Aditya, Aradhya Biswas, and Vineeth Balasubramanian. "A crowdsourced approach to student engagement recognition in e-learning environments." In 2016 IEEE winter conference on applications of computer vision (WACV), pp. 1-9. IEEE, 2016.

[7] Gupta, Abhay, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. "Daisee: Towards user engagement recognition in the wild." arXiv preprint arXiv:1609.01885 (2016).

[8] Lugaresi, Camillo, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang et al. "Mediapipe: A framework for building perception pipelines." arXiv preprint arXiv:1906.08172 (2019).

[9] Baltrušaitis, Tadas, Peter Robinson, and Louis-Philippe Morency. "Openface: an open source facial behavior analysis toolkit." In 2016 IEEE winter conference on applications of computer vision (WACV), pp. 1-10. IEEE, 2016.

[10] Tan, Xingwei, Mahathi Parvatham, Chiara Gambi, and Gabriele Pergola. "IntrEx: A Dataset for Modeling Engagement in Educational Conversations." arXiv preprint arXiv:2509.06652 (2025).

[11] Shankar, Rohan. "A multimodal approach for real-time engagement monitoring in e-learning using machine learning." In 2024 IEEE Frontiers in Education Conference (FIE), pp. 1-8. IEEE, 2024.

[12] Das, Riju, and Soumyabrata Dev. "Optimizing student engagement detection using facial and behavioral features." Neural Computing and Applications 37, no. 23 (2025): 19063-19085.

[13] Miao, Qin, Lemin Li, and Dongming Wu. "An English video teaching classroom attention evaluation model incorporating multimodal information." Journal of Ambient Intelligence and Humanized Computing 15, no. 7 (2024): 3067-3079.

[14] Nuha, Nishat Sultana, Tanjim Mahmud, Noortaz Rezaoana, Mohammad Shahadat Hossain, and Karl Andersson. "An approach of analyzing classroom student engagement in multimodal environment by using deep learning." In 2023 IEEE 9th International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), pp. 286-291. IEEE, 2023.

[15] Singh, Ruchi, Ramanujam E, and Naresh Babu M. "MMSAD—A multi-modal student attentiveness detection in smart education using facial features and landmarks." Journal of Ambient Intelligence and Smart Environments 17, no. 3 (2025): 326-348.

[16] Moise, Gabriela, Elia Georgiana Dragomir, Daniela Șchiopu, and Lidia Angelica Iancu. "Towards integrating automatic emotion recognition in education: A deep learning model based on 5 EEG channels." International Journal of Computational Intelligence Systems 17, no. 1 (2024): 230.

[17] Liao, Chen-Hsuan, and Jiun-Yu Wu. "Deploying multimodal learning analytics models to explore the impact of digital distraction and peer learning on student performance." Computers & Education 190 (2022): 104599.

[18] Acosta, Halim, Seung Lee, Bradford Mott, Haesol Bae, Krista Glazewski, Cindy Hmelo-Silver, and James Lester. "Multimodal learning analytics for predicting student collaboration satisfaction in collaborative game-based learning."

International Conference on Educational Data Mining, 2024.

[19] Yan, Lijuan, Xiaotao Wu, and Yi Wang. "Student engagement assessment using multimodal deep learning." PLoS One 20, no. 6 (2025): e0325377.

[20] Dwivedi, Deep, Ritik Garg, Shiva Baghel, Rushil Thareja, Ritvik Kulshrestha, Mukesh Mohania, and Jainendra Shukla. "Effecti-net: a multimodal framework and database for educational content effectiveness analysis." In Proceedings of the 14th learning analytics and knowledge conference, pp. 667-677. 2024.

[21] Moon, Jewoong, Fengfeng Ke, Zlatko Sokolikj, and Ibrahim Dahlstrom-Hakki. "Multimodal Data Fusion to Track Students' Distress during Educational Gameplay." Journal of Learning Analytics 9, no. 3 (2022): 75-87.

[22] Chan, Rosanna Yuen-Yan, Chun Man Victor Wong, and Yen Na Yum. "Predicting behavior change in students with special education needs using multimodal learning analytics." IEEE Access 11 (2023): 63238-63251.

[23] Zhang, T., H. A. B. Dahlan, Z. Xie, J. Wu, Y. Chen, Q. Pan, and Y. Huang. "Multimodal biometric fusion sentiment analysis of teachers and students based on classroom surveillance video streaming." Appl. Math. Nonlinear Sci 9, no. 1 (2024).