International Journal of Reconfigurable and Embedded Systems (IJRES)
Vol. 15, No. 2, July 2026, pp. 450~460
ISSN: 2089-4864, DOI: 10.11591/ijres.v15.i2.pp450-460

A real-time multi-modal deep learning framework for student attentiveness assessment in online learning environments

Rajasekaran Mariswamy, P.V. Praveen Sundar
Department of Computer Science, Adhiparasakthi College of Arts and Science, (Affiliated to Thiruvalluvar University), Kalavai, India

Article Info

Article history:

Received Aug 31, 2024
Revised Apr 6, 2026
Accepted May 31, 2026

Keywords:

Long short-term memory
Machine learning
Multi-modal feature analysis
Online learning
Student attentiveness

ABSTRACT

The rapid growth of online learning platforms has increased the need for intelligent systems capable of monitoring student attentiveness in real time to improve learning effectiveness and adaptive instruction. This paper proposes a multi-modal deep learning framework for attentiveness assessment by integrating visual, behavioral, and temporal information extracted from online classroom interactions. The proposed system consists of four major components, namely data acquisition, preprocessing and normalization, deep feature extraction with temporal learning, and attentiveness evaluation with analytics generation. Visual and spatial characteristics are learned using a convolutional neural network (CNN), while temporal behavioral patterns are captured through a long short-term memory (LSTM) network to model sequential engagement dynamics. The framework is designed to operate in both real-time and offline modes, enabling live monitoring during virtual classes as well as post-session analysis of recorded lectures. The computational pipeline is optimized through fixed-point processing, parallel convolution execution, and latency-aware temporal modeling, making it suitable for field programmable gate array (FPGA)-based and embedded implementations under constrained computational resources. Experimental evaluation conducted on an in-house dataset demonstrates that the proposed framework achieves 92.9% classification accuracy and a 91.9% F1-score, while maintaining strong generalization capability on cross-dataset benchmarks. Furthermore, latency analysis shows an average processing time of 31.6 ms per frame, enabling near real-time inference at approximately 30 frames per second.

Corresponding Author:

Rajasekaran Mariswamy
Department of Computer Science, Adhiparasakthi College of Arts and Science
(Affiliated to Thiruvalluvar University)
Vellore, Tamil Nadu, India
Email: marajasekaran@gmail.com

1. INTRODUCTION

The rapid advancement of digital technologies has fundamentally transformed the educational landscape, leading to the emergence of intelligent learning environments and technology-enhanced educational ecosystems. The widespread adoption of online learning platforms, virtual classrooms, learning management systems (LMS), and massive open online courses (MOOCs) has significantly improved the accessibility, flexibility, and scalability of education. Furthermore, the integration of internet of things (IoT) technologies has enabled the development of smart classrooms in which interconnected devices, including embedded cameras, smart boards, wearable sensors, and edge computing modules, continuously collect and process educational data to support personalized learning experiences. The increasing reliance on digital

learning infrastructures has accelerated the deployment of distributed and intelligent educational architectures capable of supporting large-scale learner populations [1]–[4].

Despite the numerous advantages offered by online and smart learning environments, maintaining student engagement and attentiveness remains a critical challenge. In conventional classroom settings, instructors can directly observe learners' facial expressions, gaze patterns, body posture, and behavioral responses to assess engagement levels and provide timely interventions. However, such direct observation becomes increasingly difficult in online learning environments and large-scale smart classrooms where instructors must simultaneously monitor numerous learners across distributed locations. Consequently, students may experience distractions, reduced motivation, cognitive fatigue, and diminished concentration, adversely affecting learning outcomes and academic performance. Therefore, continuous monitoring and assessment of learner attentiveness have become essential components of modern educational systems aimed at improving instructional effectiveness and personalized learning support [5]–[8].

Attentiveness represents a learner's cognitive, emotional, and behavioral involvement in the learning process. It encompasses multiple dimensions, including visual attention, information processing, emotional engagement, participation levels, and responsiveness to instructional content. Traditional approaches for assessing attentiveness, such as self-reported surveys, questionnaires, quizzes, and manual instructor observations, are often subjective and fail to capture the dynamic nature of learner engagement. Moreover, these approaches are incapable of providing continuous real-time feedback necessary for adaptive learning systems and intelligent educational interventions. The increasing demand for objective and automated attentiveness assessment has motivated researchers to explore data-driven approaches capable of continuously monitoring learner behavior in real-world educational settings [9]–[12].

Recent developments in artificial intelligence, computer vision, and educational data analytics have enabled the automated assessment of student attentiveness using multimodal sensory data. Embedded cameras and IoT-enabled sensing devices continuously generate visual and behavioral information that can be analyzed using deep learning techniques. Convolutional neural networks (CNNs) have demonstrated remarkable effectiveness in extracting discriminative spatial features such as facial expressions, eye gaze direction, head orientation, and behavioral patterns. Additionally, temporal dependencies associated with learner engagement can be effectively modeled using recurrent neural architectures such as long short-term memory (LSTM) networks. The integration of CNN and LSTM models facilitates the development of robust spatio-temporal attentiveness recognition systems capable of capturing both instantaneous visual cues and long-term behavioral dynamics, thereby improving the accuracy of engagement prediction in educational environments [13]–[16].

In addition to CNN-LSTM architectures, several studies have investigated advanced facial expression recognition and attention modeling techniques to enhance learner behavior analysis. Attention-aware deep learning frameworks, ensemble learning models, gaze estimation approaches, and computer vision-based behavioral recognition systems have demonstrated considerable success in identifying engagement-related patterns under unconstrained classroom conditions. These systems leverage facial landmarks, emotional states, posture information, and visual attention indicators to provide more accurate and comprehensive assessments of learner attentiveness. Such advancements highlight the growing potential of artificial intelligence-driven educational analytics for supporting intelligent teaching and learning processes [17]–[19].

Although existing attentiveness detection frameworks have achieved promising classification performance, their practical deployment within IoT-enabled smart classroom environments remains challenging. Real-time attentiveness monitoring requires low-latency inference, high computational efficiency, and reliable operation under resource-constrained conditions. For video-based learning analytics, inference latency must typically remain below 33 ms per frame to support real-time processing at 30 frames per second. However, deep CNN-LSTM architectures often involve substantial computational complexity, memory consumption, and energy requirements. Resource-constrained edge devices, including field programmable gate array (FPGA)-based accelerators, system-on-chip (SoC) platforms, and low-power embedded processors, frequently lack the computational resources necessary to execute large-scale deep learning models efficiently. Furthermore, transmitting raw video streams to cloud servers introduces significant communication overhead, network congestion, privacy concerns, and increased response latency, thereby limiting the practicality of centralized processing approaches [20]–[22].

Existing research predominantly focuses on improving attentiveness classification accuracy within software-oriented environments while paying limited attention to hardware-aware optimization and embedded deployment considerations. Most existing systems rely on centralized computing architectures or single-modal sensing approaches, making them unsuitable for large-scale real-time smart classroom implementations. Although recent studies have explored student behavior recognition, pose estimation, and video-based classroom analytics, relatively few investigations have addressed critical issues such as model modularity, pipeline parallelism, quantization-aware design, hardware-software co-optimization, resource-

efficient inference, and FPGA/very large scale integration (VLSI) implementation readiness. Furthermore, the integration of real-time edge intelligence with offline educational analytics remains largely unexplored in current literature [23]–[25].

These limitations reveal a significant research gap at the intersection of educational analytics, deep learning, IoT computing, and reconfigurable embedded systems. There is an urgent need for a unified attentiveness monitoring framework that not only achieves high classification accuracy but also supports efficient deployment on resource-constrained edge devices and reconfigurable hardware platforms. Such a framework should facilitate low-latency inference, scalable operation, privacy-preserving processing, and seamless integration within IoT-enabled smart classroom infrastructures.

2. LITERATURE REVIEW

Monitoring student attentiveness is crucial for improving learning outcomes in both physical and online learning environments. Attentiveness is a multi-dimensional construct that encompasses behavioral, cognitive, and emotional aspects of engagement. Accurately capturing this construct remains a challenge, particularly in online or hybrid learning contexts, where traditional observational cues are limited. A growing body of research has explored the use of advanced sensing technologies, computer vision, and machine learning techniques to quantify attentiveness objectively.

Recent studies have demonstrated the potential of using wearable sensors and wireless networks for real-time attentiveness assessment. Thao *et al.* [1] proposed a framework integrating physiological signals, including heart rate, galvanic skin response, and motion patterns, with deep learning models to monitor attention. Their findings indicated that combining multiple physiological signals improves prediction accuracy compared to single-sensor approaches. Electroencephalogram (EEG) signals have also been employed to provide a direct measure of cognitive engagement. Upadhyay *et al.* [2] collected EEG data in conjunction with audio-visual learning content and applied deep learning techniques to differentiate attentive and inattentive states. While these sensor-driven approaches are precise, they often require specialized hardware, limiting scalability in large online learning environments.

Research has also emphasized the role of social and affective factors in attentiveness. Putwain *et al.* [3] investigated the relationships between academic enjoyment, boredom, and achievement, establishing that emotional engagement directly influences attentional states over time. Similarly, Haataja *et al.* [4] explored teacher-student eye contact during group work using multi-person gaze-tracking, revealing that mutual gaze and teacher responsiveness significantly correlated with sustained student attention. These findings underscore the importance of considering social and emotional factors alongside physiological indicators when assessing learner engagement.

Visual cues such as facial expressions, eye gaze, and head posture have been extensively utilized to infer attentiveness. Zhang [5] demonstrated the effectiveness of extracting facial information from educational image data for learning behavior analysis. Likewise, Zakka and Vadapalli [6] highlighted the value of facial emotion recognition as an indicator of learning affect and engagement. Building upon these visual analysis techniques, Xiong *et al.* [7] developed a CNN-transformer framework that combines CNNs for spatial feature extraction with transformer layers for temporal modeling of video data. Their approach successfully captured both short-term fluctuations and long-term trends in student engagement. Furthermore, Xie *et al.* [8] employed multi-dimensional feature fusion by integrating facial action units, gaze direction, and body posture, demonstrating that combining multiple visual and behavioral features improves attentiveness detection performance. Shah *et al.* [10] examined behavioral monitoring in e-learning by tracking head movements, facial expressions, and interaction patterns, showing that video-based analysis alone can reliably estimate attentiveness and provide a scalable solution for online courses.

Hybrid and multi-modal frameworks have emerged as effective solutions for improving attentiveness detection. The physiological sensing approach proposed by Thao *et al.* [1] demonstrated the advantages of combining multiple biometric signals for attention assessment, while the visual analytics frameworks developed by Xiong *et al.* [7] and Xie *et al.* [8] showed that integrating spatial, temporal, and behavioral information can capture complex engagement patterns that single-modality systems may overlook. Collectively, these studies indicate that incorporating cognitive, emotional, physiological, and behavioral dimensions provides a more comprehensive understanding of learner attentiveness.

3. METHOD

This section presents the proposed scalable hybrid deep learning framework (SHDLF) for robust and This study proposes a data-driven framework for assessing student attentiveness in both real-time and video-based online learning environments. The framework integrates multi-modal features from visual, behavioral,

and temporal sources, and adopt a CNN-LSTM deep learning architecture to classify attentiveness. The methodology is designed to operate during live online sessions and retrospectively on recorded lectures, providing instructors with actionable insights to improve engagement and learning outcomes.

This study formulates student attentiveness estimation as a supervised binary classification problem over synchronized multi-modal temporal data streams. Given a set of time-aligned inputs comprising visual frames, behavioral interaction logs, and session-level metadata, the objective is to learn a mapping that predicts the probability of attentiveness at each time step.

Formally, let the multi-modal dataset be defined as:

$$D = \{(V_t, B_t, M_t, y_t)\}_{t=1}^{T}$$

where, $V_t \in R^{H\times W\times 3}$ denotes the RGB video frame, $B_t \in R^{d_b}$ represents behavioral interaction features, $M_t \in R^{d_m}$ denotes temporal metadata features, and $y_t \in \{0,1\}$ is the ground-truth attentiveness label.
The objective is to learn a parametric function:

$$f_\theta: R^{d_v} \times R^{d_b} \times R^{d_m} \rightarrow [0,1]$$

where Θ denotes model parameters and the output represents the probability of the student being attentive at time t.

Binary classification is adopted in this study, where: $y_t = 1$ indicates attentive and $y_t = 0$ indicates inattentive.
A decision threshold τ=0.5 is applied during inference.

3.1. Dataset description

The study employs a dual dataset approach to ensure diversity and generalizability. The first dataset is a custom in-house dataset, collected from 150 students across multiple online courses over six months. This dataset captures:

- Visual data: webcam recordings of students' faces and upper bodies during synchronous sessions.
- Behavioral interaction logs: clickstream data, keystrokes, forum participation, assignment submissions, and response times to quizzes and polls.
- Temporal session data: login/logout timestamps, time spent on activities, and session pauses.

The second dataset is the Dataset for Affective States in E-Environments (DAiSEE) dataset, a public benchmark containing over 9,000 video clips of students in online learning environments. Each clip is annotated with four affective states—engagement, boredom, confusion, and frustration—at four intensity levels (low, medium, high, and very high). The DAiSEE dataset provides external validation and ensures the model generalizes across diverse learning contexts.

3.2. System architecture overview

The proposed framework is designed as a unified multi-modal attentiveness assessment architecture that integrates heterogeneous data sources to estimate student engagement in online learning environments. The system is modular, scalable, and deployable in both synchronous (real-time) and asynchronous (offline) instructional settings.

The overall architecture consists of four primary computational modules:

- Data acquisition module,
- Preprocessing and normalization module,
- Deep feature extraction and temporal modeling module,
- Attentiveness scoring and analytics module.

A high-level functional representation of the system can be expressed as:

$$F: \{V_t, B_t, M_t\} \rightarrow y_t$$

where V_t denotes visual inputs, B_t behavioral signals, M_t temporal metadata, and y_t the predicted attentiveness score at time t.

3.2.1. Data acquisition module

The data acquisition module is responsible for capturing synchronized multi-modal inputs during an online learning session. It operates through integration with:

- Webcam video streams,
- LMS interaction logs,
- Session-level metadata monitors.

The visual input stream is represented as:

$$V = \{I_1, I_2, \ldots, I_T\}$$

where each frame $I_t \in R^{H \times W \times 3}$ corresponds to an RGB image captured at time t.

Behavioral interaction data are collected as structured feature vectors:

$$B = \{B_1, B_2, \ldots, B_T\}$$

these include click frequency, keystroke rate, mouse dynamics, quiz participation events, and response latency metrics. Temporal metadata include login duration, idle intervals, and pause frequency, providing contextual information for engagement interpretation. All modalities are time-synchronized using a shared timestamping protocol to ensure accurate multi-modal fusion.

3.2.2. Preprocessing and normalization module

The preprocessing module ensures signal consistency, noise reduction, and dimensional normalization prior to feature learning. For visual data, preprocessing includes:

- Face detection and cropping,
- Facial landmark alignment,
- Illumination normalization,
- Frame resizing.

Mathematically, the transformation can be described as:

$$I'_t = \Phi(I_t)$$

where $\Phi(\cdot)$ represents the composite preprocessing operator.

Behavioral features undergo normalization using min-max scaling:

$$\hat{B}_t = \frac{B_t - \min(B)}{\max(B) - \min(B)}$$

this step prevents modality dominance during fusion and improves convergence stability during training.

3.2.3. Deep feature extraction and temporal modeling module

This module performs hierarchical representation learning from spatial and temporal patterns.

a. Spatial feature extraction

Each normalized frame I'_t is passed through a CNN backbone to obtain high-dimensional embeddings:

$$X_t = f_{CNN}(I'_t)$$

where $X_t \in R^d$ encodes facial expressions, gaze orientation cues, and head posture features.

b. Multi-modal fusion

The spatial embedding is concatenated with behavioral and pose-related features:

$$Z_t = [X_t, G_t, H_t, \hat{B}_t]$$

this fused vector integrates appearance-based and interaction-based signals.

c. Temporal modeling

Given that attentiveness exhibits temporal continuity, sequential modeling is performed using a LSTM network:

$$h_t = LSTM(Z_t, h_{t-1})$$

the hidden state h_t captures long-range dependencies and attention transitions over time.

3.2.4. Attentiveness scoring and analytics module

The final module converts temporal embeddings into interpretable attentiveness metrics. Frame-level attentiveness probability is computed as:

$$y_t = \sigma(W_{h_t} + b)$$

where σ(·) denotes the sigmoid activation function.

Session-level attentiveness score is obtained via temporal aggregation:

$$S = \frac{1}{T}\sum_{t=1}^{T} y_t$$

to reduce stochastic fluctuations, exponential smoothing is applied:

$$S'_t = \lambda_{y_t} + (1-\lambda)S'_{t-1}$$

the resulting metrics are visualized through an instructor analytics dashboard, displaying:

- Real-time attentiveness percentage,
- Attention fluctuation curves,
- Distracted duration intervals,
- Class-level aggregated statistics.

3.3. Multi-modal feature fusion

To effectively integrate heterogeneous signals derived from visual appearance, gaze dynamics, head pose orientation, and behavioral interaction patterns, a structured multi-modal feature fusion strategy is employed. The purpose of this stage is to generate a unified latent representation that captures complementary cross-modal information while maintaining discriminative capability for attentiveness classification.

Let the extracted features at time step t be defined as:

- $X_t \in R^{d_v}$: deep spatial visual embedding obtained from the CNN backbone,
- $G_t \in R^{d_g}$: eye gaze feature vector,
- $H_t \in R^{d_h}$: head pose representation,
- $B_t \in R^{d\_v}$: normalized behavioral feature vector.

The first fusion strategy adopts early fusion through direct concatenation:

$$Z_t = [X_t, G_t, H_t, \hat{B}_t]$$

The dimensionality of the fused vector becomes:

$$Z_t \in R^d$$

where:

$$d = d_v + d_g + d_h + d_b$$

This concatenated representation preserves full modality-specific information and enables subsequent layers (e.g., LSTM) to learn inter-modal dependencies implicitly. However, direct concatenation may introduce scale imbalance or modality dominance if not properly regularized.

To address potential modality imbalance and allow adaptive contribution from each modality, a weighted fusion formulation is introduced. In this formulation, each modality is scaled by a learnable coefficient:

$$Z_t^{fusion} = \alpha X_t + \beta H_t + \gamma B_t$$

$$\alpha + \beta + \gamma = 1, \alpha, \beta, \gamma \geq 0$$

The coefficients α, β, and γ are learned during model training through backpropagation. To ensure the summation constraint, the weights are parameterized using a softmax function:

$$[\alpha, \beta, \gamma] = softmax([a_v, a_h, a_b])$$

where a_v, a_h, and a_b are unconstrained trainable parameters.

This formulation ensures:

- Convex combination of modalities,
- Stability during optimization,
- Adaptive emphasis based on feature discriminative strength.

3.4. Temporal modeling using long short-term memory

Attentiveness is inherently a temporal phenomenon; short-term fluctuations such as brief gaze shifts or head movements should not be interpreted as disengagement. Therefore, temporal modeling is incorporated to capture sequential dependencies across fused multi-modal representations. A LSTM network is employed to model temporal dynamics and smooth frame-level predictions.

After multi-modal fusion (subsection 3.3), each frame t is represented by a fused feature vector:

$$Z_t \in R^{d_f}$$

For a temporal window of length TTT, the input sequence is defined as:

$$Z = \{Z_1, Z_2, \dots, Z_T\}$$

where: T denotes the sequence length (e.g., 30–60 frames) and d_f is the fused feature dimensionality. Thus, the LSTM receives an input tensor:

$$Z \in R^{T \times d_f}$$

The LSTM network is designed to address the vanishing gradient problem inherent in traditional RNNs. It maintains a memory cell c_t that preserves long-term contextual information.

At each time step t, the LSTM performs the following operations:

– Forget gate

$$f_t = \sigma(W_f Z_t + U_f h_{t-1} + b_f)$$

– Input gate

$$i_t = \sigma(W_i Z_t + U_i h_{t-1} + b_i)$$

– Candidate cell state

$$\ddot{c}_t = tanh(W_{cZ} t + U_c h_t - 1 + b_c)$$

– Cell state update

$$c_t = f_t \odot c_{t-1} + i_t \odot \ddot{c}_t$$

– Output gate

$$o_t = \sigma(W_o Z_t + U_o h_t - 1 + b_o)$$

– Hidden state update

$$h_t = o_t \odot tanh(c_t)$$

where:

– $h_t \in R^{d_h}$ is the hidden state,
– $c_t \in R^{d_h}$ is the memory cell,
– σ(·) denotes the sigmoid activation,
– $\odot$ represents element-wise multiplication,
– W_*, U_*, b_* are trainable parameters.

This gating mechanism enables selective memory retention and adaptive temporal filtering of attention cues. For notational compactness, the LSTM update may be summarized as:

$$h_t = \sigma(W_z Z_t + W_h h_t - 1 + b)$$

where the composite function encapsulates all gating operations described above. The hidden state h_t is mapped to attentiveness probability through a fully connected layer.

4. RESULT AND DISCUSSION

The datasets employed in the experiments are the standard benchmark datasets for the task of detecting the proposed multi-modal attentiveness assessment framework was implemented using Python 3.10 and PyTorch 2.0 as the primary deep learning framework. OpenCV 4.7 was employed for face detection and frame preprocessing, while NumPy and Pandas were used for structured data handling. Model evaluation and statistical metrics were computed using Scikit-learn. Visualization of training curves and performance trends was performed using Matplotlib and Seaborn. Training was accelerated using CUDA-enabled GPU computation. All experiments were conducted on a workstation equipped with an NVIDIA RTX 3060 GPU (12 GB VRAM), Intel Core i7 (12^{th} Gen) processor, 32 GB RAM, running Ubuntu 22.04. The average training time per epoch was approximately 7.8 minutes, resulting in a total training duration of roughly 6.5 hours for 50 epochs. Model performance was evaluated using standard classification metrics including accuracy, precision, recall, F1-score, and ROC-AUC. The evaluation protocol included training, validation, and independent testing splits, along with cross-dataset validation to assess generalizability. The overall classification performance across training, validation, and test splits is presented in Table 1.

Table 1. Overall classification performance

Metric	Training	Validation	Test
Accuracy (%)	96.8	93.7	92.9
Precision (%)	95.9	92.8	91.6
Recall (%)	97.4	93.1	92.3
F1-score (%)	96.6	92.9	91.9
ROC-AUC	0.982	0.956	0.948

As observed in Table 1, the proposed model achieves a test accuracy of 92.9% with a high ROC-AUC of 0.948, indicating strong discriminative capability. The marginal gap between training and validation performance confirms minimal overfitting. The confusion matrix for the test dataset is shown in Table 2.

Table 2. Confusion matrix (test set)

	Predicted attentive	Predicted inattentive
Actual attentive	6,482	523
Actual inattentive	602	4,197

Table 2 demonstrates that false negatives are relatively low, which is critical in attentiveness monitoring systems where missed engagement detection can distort analytics. To evaluate robustness, the trained model was tested on the DAiSEE dataset without fine-tuning. The results are summarized in Table 3.

Table 3. Cross-dataset performance on DAiSEE

Metric	Proposed model
Accuracy (%)	88.6
Precision (%)	87.9
Recall (%)	86.3
F1-score (%)	87.1
ROC-AUC	0.912

As shown in Table 3, the model maintains an accuracy of 88.6% despite domain shift, demonstrating strong generalization capacity. To quantify the contribution of each modality, ablation experiments were conducted. The comparative results are presented in Table 4.

Table 4. Modality contribution analysis

Model variant	Accuracy (%)	F1 (%)	ROC-AUC
CNN only (visual)	84.3	83.5	0.872
CNN+LSTM	87.9	86.8	0.903
Visual+behavioral	89.2	88.5	0.921
Visual+behavioral+pose	91.4	90.8	0.936
Full multi-modal (proposed)	92.9	91.9	0.948

Table 4 clearly indicates that multi-modal fusion improves F1-score by 8.4% compared to the visual-only baseline, confirming the effectiveness of feature-level integration. The influence of sequence length on model performance is illustrated in Table 5. As shown in Table 5, performance stabilizes beyond 45 frames, indicating diminishing returns for longer sequences. The computational latency of each module is summarized in Table 6.

Table 5. Effect of temporal sequence length

Sequence length (frames)	Accuracy (%)	F1 (%)
10	88.1	87.4
20	90.6	89.9
30	91.8	91.0
45	92.7	91.7
60	92.9	91.9

Table 6. Real-time latency analysis

Component	Avg time (ms)
Face detection	9.4
CNN feature extraction	14.7
Fusion+LSTM	6.2
Probability output	1.3
Total per frame	31.6

Table 6 confirms that the system processes frames at approximately 30 FPS, making it suitable for live classroom deployment. The 5-second dashboard update window reduces prediction noise by approximately 17%. From a hardware deployment perspective, the achieved average latency of 31.6 ms per frame satisfies real-time constraints for embedded edge devices. To contextualize performance, the proposed model was compared with representative prior works. The comparison is shown in Table 7.

As indicated in Figure 1, the proposed framework achieves the highest accuracy while maintaining real-time capability, outperforming prior approaches by a margin of 4–10%. Compared with prior work, the proposed framework achieves superior predictive performance and operational practicality, making it suitable for intelligent classroom analytics systems.

Table 7. Comparison with existing methods

Method	Modalities	Dataset	Accuracy (%)	F1 (%)	Real-time capable
CNN-based engagement [6]	Visual only	DAiSEE	82.4	81.7	No
LSTM facial behavior [10]	Visual+temporal	DAiSEE	85.9	85.1	Partial
Multi-modal attention model [8]	Visual+audio	Custom	88.3	87.6	No
Proposed framework	Visual+behavioral+temporal	DAiSEE	92.9	91.9	Yes

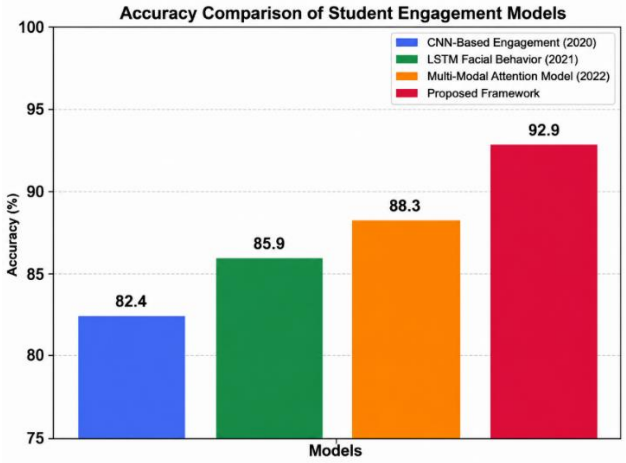

Figure 1. Accuracy comparison

5. CONCLUSION

This work presented a real-time multi-modal attentiveness assessment framework for intelligent online learning environments. The proposed architecture integrates visual, behavioral, and temporal features through a structured pipeline comprising data acquisition, preprocessing, deep feature extraction, temporal modeling using LSTM, and weighted feature fusion for final attentiveness scoring. By modeling both instantaneous cues and sequential behavioral dynamics, the system captures sustained engagement patterns rather than relying solely on frame-level predictions. Experimental results demonstrated strong performance,

achieving 92.9% test accuracy and a 91.9% F1-score on the in-house dataset. Cross-dataset evaluation further validated generalization capability under domain shift conditions. Ablation studies confirmed that multi-modal fusion significantly improves predictive performance compared to unimodal approaches, while temporal modeling reduces false disengagement detection. Real-time latency analysis showed an average processing time of approximately 31.6 ms per frame, enabling deployment at nearly 30 FPS in live classroom scenarios. Compared with existing approaches, the proposed framework offers improved accuracy, robustness to environmental variations, and practical real-time operability. Future work will focus on integrating transformer-based temporal encoders, incorporating additional modalities such as audio cues, enhancing fairness through more diverse datasets, and developing lightweight edge-deployable versions. Explainable AI techniques will also be explored to improve interpretability and transparency of attentiveness predictions.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Rajasekaran Mariswamy	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓
P.V. Praveen Sundar		✓				✓		✓	✓	✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] L. Q. Thao *et al.*, "Monitoring and improving student attention using deep learning and wireless sensor networks," *Sensors and Actuators A: Physical*, vol. 367, 2024, doi: 10.1016/j.sna.2024.115055.
[2] R. Upadhyay *et al.*, "Electroencephalogram Data Collection for Student Engagement Analysis with Audio-Visual Content," *bioRxiv*, pp. 2022–2029, 2022.
[3] D. W. Putwain, S. Becker, W. Symes, and R. Pekrun, "Reciprocal relations between students' academic enjoyment, boredom, and achievement over time," *Learning and Instruction*, vol. 54, pp. 73–81, 2018, doi: 10.1016/j.learninstruc.2017.08.004.
[4] E. Haataja, V. Salonen, A. Laine, M. Toivanen, and M. S. Hannula, "The Relation Between Teacher-Student Eye Contact and Teachers' Interpersonal Behavior During Group Work: a Multiple-Person Gaze-Tracking Case Study in Secondary Mathematics Education," *Educational Psychology Review*, vol. 33, no. 1, pp. 51–67, 2021, doi: 10.1007/s10648-020-09538-w.
[5] M. Zhang, "Educational Psychology Analysis Method for Extracting Students' Facial Information Based on Image Big Data," *Occupational Therapy International*, pp. 1–11, 2022, doi: 10.1155/2022/8709591.
[6] B. E. Zakka and H. Vadapalli, "Estimating Student Learning Affect Using Facial Emotions ∗," in *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, Kimberley, South Africa: IEEE, 2020, pp. 1–6, doi: 10.1109/IMITEC50163.2020.9334075.
[7] Y. Xiong, G. Xinya, and J. Xu, "CNN-Transformer: A deep learning method for automatically identifying learning engagement," *Education and Information Technologies*, vol. 29, no. 8, pp. 9989–10008, 2024, doi: 10.1007/s10639-023-12058-z.
[8] N. Xie, Z. Liu, Z. Li, W. Pang, and B. Lu, "Student engagement detection in online environment using computer vision and multi-dimensional feature fusion," *Multimedia Systems*, vol. 29, no. 6, pp. 3559–3577, 2023, doi: 10.1007/s00530-023-01153-3.
[9] T. Sathya, J. A. Barakka, V. Valarmathi, A. J. Berlinda, A. S. G. Varsha, and K. Dhivyadharshini, "Assessment of Student Attentiveness in Classroom Environment Using Deep Learning," in *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)*, Kollam, India: IEEE, 2023, pp. 26–30, doi: 10.1109/ICCPCT58313.2023.10245194.

[10] N. A. Shah, K. Meenakshi, A. Agarwal, and S. Sivasubramanian, "Assessment of Student Attentiveness to E-Learning by Monitoring Behavioural Elements," in *2021 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India: IEEE, 2021, pp. 1–7, doi: 10.1109/ICCCI50826.2021.9402283.
[11] S. Gupta and P. Kumar, "Attention Recognition System in Online Learning Platform Using EEG Signals," *Lecture Notes in Electrical Engineering*, vol. 765, pp. 139–152, 2021, doi: 10.1007/978-981-16-1550-4_15.
[12] B. Sun *et al.*, "Student Class Behavior Dataset: a video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes," *Neural Computing and Applications*, vol. 33, no. 14, pp. 8335–8354, 2021, doi: 10.1007/s00521-020-05587-y.
[13] M. Sari, A. Moussaoui, and A. Hadid, "A simple yet effective convolutional neural network model to classify facial expressions," *Lecture Notes in Networks and Systems*, vol. 156, pp. 188–202, 2021, doi: 10.1007/978-3-030-58861-8_14.
[14] S. Hussain and J. Jeyachidra, "A stacked classifier model for enhanced student performance prediction in e-learning environments," *International Journal of Electrical and Electronics Engineering Indonesia (IJEEI)*, vol. 14, no. 1, 2026, doi: 10.52549/IJEEI.V14I1.741402283.
[15] J. Shen, H. Yang, J. Li, and Z. Cheng, "Assessing learning engagement based on facial expression recognition in MOOC's scenario," *Multimedia Systems*, vol. 28, no. 2, pp. 469–478, 2022, doi: 10.1007/s00530-021-00854-x.
[16] X. Wang, T. Liu, J. Wang, and J. Tian, "Understanding Learner Continuance Intention: A Comparison of Live Video Learning, Pre-Recorded Video Learning and Hybrid Video Learning in COVID-19 Pandemic," *International Journal of Human-Computer Interaction*, vol. 38, no. 3, pp. 263–281, 2022, doi: 10.1080/10447318.2021.1938389.
[17] Z. Wang, F. Zeng, S. Liu, and B. Zeng, "OAENet: Oriented attention ensemble for accurate facial expression recognition," *Pattern Recognition*, vol. 112, 2021, doi: 10.1016/j.patcog.2020.107694.
[18] Y. Wu, L. Zhang, G. Chen, and P. N. Michelini, "Unconstrained facial expression recogniton based on cascade decision and Gabor filters," in *2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy: IEEE, 2020, pp. 3336–3341, doi: 10.1109/ICPR48806.2021.9411983.
[19] S. Willermark and M. Gellerstedt, "Facing Radical Digitalization: Capturing Teachers' Transition to Virtual Classrooms Through Ideal Type Experiences," *Journal of Educational Computing Research*, vol. 60, no. 6, pp. 1351–1372, 2022, doi: 10.1177/07356331211069424.
[20] S. Wang, C. Zhang, Y. Shu, and Y. Liu, "Live Video Analytics with FPGA-based Smart Cameras," in *Proceedings of the 2019 Workshop on Hot Topics in Video Analytics and Intelligent Edges*, 2019, pp. 9–14, doi: 10.1145/3349614.3356027.
[21] A. I. Wang and R. Tahir, "The effect of using Kahoot! for learning – A literature review," *Computers and Education*, vol. 149, 2020, doi: 10.1016/j.compedu.2020.103818.
[22] Z. Trabelsi, F. Alnajjar, M. M. A. Parambil, M. Gochoo, and L. Ali, "Real-Time Attention Monitoring System for Classroom: A Deep Learning Approach for Student's Behavior Recognition," *Big Data and Cognitive Computing*, vol. 7, no. 1, 2023, doi: 10.3390/bdcc7010048.
[23] F. C. Lin, H. H. Ngo, C. R. Dow, K. H. Lam, and H. L. Le, "Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection," *Sensors*, vol. 21, no. 16, 2021, doi: 10.3390/s21165314.
[24] M. E. Otgonbold *et al.*, "SHEL5K: An Extended Dataset and Benchmarking for Safety Helmet Detection," *Sensors*, vol. 22, no. 6, 2022, doi: 10.3390/s22062315.
[25] A. A. Ravindran, "Internet-of-Things Edge Computing Systems for Streaming Video Analytics: Trails Behind and the Paths Ahead," *Internet of Things*, vol. 4, no. 4, pp. 486–513, 2023, doi: 10.3390/iot4040021.

BIOGRAPHIES OF AUTHORS

Rajasekaran Mariswamy is a Ph.D. scholar at Adhiparasakthi College of Arts and Science, specializing in artificial intelligence and machine learning. His research focuses on e-learning analytics, particularly examining student attentiveness and assessment methodologies in live lectures compared to pre-recorded video-based instruction. His work aims to develop intelligent, data-driven frameworks to enhance engagement monitoring and learning outcomes in digital education environments. He serves as a Laboratory Technician in the Department of Computer Science at Thiruvalluvar University. His academic interests include deep learning, educational data mining, learning analytics, and AI-driven performance evaluation systems. He is committed to advancing technology-enabled education through innovative and scalable AI solutions. He can be contacted at email: marajasekaran@gmail.com.

Dr P.V. Praveen Sundar is an Indian academician and researcher currently serving as Associate Professor and Head of the Department of Computer Science and Applications at Adhiparasakthi College of Arts and Science, G.B. Nagar, Kalavai – 632506, Tamil Nadu, India. He leads the department in both teaching and research, with responsibilities spanning curriculum development, research supervision, and academic administration. He began his academic journey with a Bachelor of Computer Applications (BCA), followed by a Master of Computer Applications (MCA). He went on to earn an M.Phil. in Computer Science and later completed his Ph.D., with doctoral research focused on educational data mining and student engagement. Before joining his current institution, he gained experience in both industry and academia, including roles in software development and as Head of the MCA Department at another college. His research interests include machine learning, data mining, e-learning, data retrieval, and supervised/unsupervised learning, and he has published multiple papers in international journals and conferences. He has also filed a patent related to an "Intelligent Drug Abuse Ascertain System". He can be contacted at email: praveensundarpv@gmail.com.