International Journal of Informatics and Communication Technology (IJ-ICT)
Vol. 15, No. 3, September 2026, pp. 935~943
ISSN: 2252-8776, DOI: 10.11591/ijict.v15i3.pp935-943

Cross-modal attention fusion using vision transformers for robust student attentiveness estimation

Rajasekaran Mariswamy, Praveen Sundar
Department of Computer Science, Adhiparasakthi College of Arts and Science, Thiruvalluvar University, Vellore, India

Article Info

Article history:

Received Jun 21, 2024
Revised Apr 22, 2026
Accepted May 7, 2026

Keywords:

Cardiovascular disease
Deep learning
Echocardiogram imaging
IoT
Prediction
VGNet-CNN

ABSTRACT

Automated student attentiveness estimation is a fundamental component of intelligent e-learning systems and adaptive classroom analytics. Traditional convolutional and recurrent architectures often struggle to model long-range temporal dependencies and complex inter-modal relationships inherent in engagement behavior. To address these limitations, this paper proposes a cross-modal attention fusion framework built upon a vision transformer (ViT) backbone for robust student attentiveness estimation. The proposed architecture leverages patch-based visual encoding through a ViT to capture global spatial dependencies, while behavioral cues such as gaze direction, head pose, and blink dynamics are embedded into a shared latent representation space. A cross-modal multi-head attention mechanism is introduced to dynamically learn interactions between visual and behavioral modalities, replacing static weighted fusion strategies. Temporal dynamics are modeled using a Transformer encoder, enabling effective long-range sequence modeling without recurrent dependencies. Experimental evaluation on a benchmark attentiveness dataset demonstrates superior performance compared to CNN–LSTM-based models, achieving improved accuracy, F1-score, and robustness under challenging lighting and occlusion conditions. Ablation studies validate the contribution of cross-modal attention and transformer-based temporal modeling. The proposed framework maintains real-time feasibility while significantly enhancing discriminative capability.

Corresponding Author:

Rajasekaran Mariswamy
Department of Computer Science, Adhiparasakthi College of Arts and Science, Thiruvalluvar University
Vellore, India
Email: marajasekaran@gmail.com

1. INTRODUCTION

Cardiovascular diseases (CVDs) remain the leading cause of death globally, accounting for an estimated. The rapid expansion of online education and virtual classrooms has fundamentally transformed the landscape of teaching and learning. Massive open online courses (MOOCs), hybrid instructional models, and fully remote academic programs have enabled scalable access to education across geographical and socio-economic boundaries. However, this digital transformation has also introduced a critical challenge: the difficulty of monitoring student attentiveness in the absence of physical classroom supervision [1]-[3]. Unlike traditional in-person environments where instructors can observe facial expressions, posture, and behavioral cues to gauge engagement, online platforms often lack reliable mechanisms to assess whether learners are attentive, distracted, or disengaged. Consequently, automated attentiveness monitoring has emerged as a crucial component of intelligent e-learning systems, enabling adaptive content delivery, personalized feedback, and data-driven pedagogical interventions [4]-[6].

Student attentiveness is inherently a multi-dimensional construct that manifests through visual and behavioral signals such as gaze direction, head orientation, facial micro-expressions, blink rate, and posture dynamics. These signals evolve over time and interact in complex ways. Recent advances in computer vision and deep learning have facilitated the development of automated engagement detection systems, primarily leveraging convolutional neural networks (CNNs) for spatial feature extraction and recurrent neural networks (RNNs), particularly long short-term memory (LSTM) models, for temporal modeling. While these architectures have achieved moderate success, they exhibit notable limitations when applied to the nuanced task of attentiveness estimation in unconstrained online environments [7]-[9].

CNN-based approaches excel at capturing local spatial patterns through hierarchical feature extraction. However, their inherent inductive bias toward locality restricts their ability to model global contextual relationships within an image. Engagement-related cues often require holistic interpretation; for instance, subtle combinations of eye gaze deviation and head pose shifts may indicate cognitive disengagement. Standard CNN architectures rely on progressively stacked convolutional filters to enlarge the receptive field, but this process remains constrained by spatial locality and depth-dependent aggregation. Consequently, CNN-based systems may struggle to capture long-range spatial dependencies that are essential for robust attentiveness inference [10]-[12].

To address temporal dynamics, many existing works incorporate LSTM or gated recurrent unit (GRU) networks. Recurrent architectures are designed to process sequential data by maintaining hidden states that propagate information across time steps. Although LSTMs mitigate the vanishing gradient problem relative to vanilla RNNs, they still exhibit limitations in modeling long-range dependencies, particularly in extended video sequences. The sequential nature of recurrence introduces computational inefficiencies and restricts parallelization during training [13]-[17]. Moreover, the memory mechanism of LSTMs may not effectively capture complex temporal interactions spanning long intervals, especially when attentiveness patterns fluctuate gradually over time. As a result, recurrent models can produce unstable or delayed responses to subtle engagement shifts.

Two core challenges therefore emerge in attentiveness estimation: (1) effective modeling of long-range temporal dependencies, and (2) dynamic inter-modal feature interaction. First, attentiveness is not solely determined by instantaneous behavior but rather by sustained patterns across extended temporal windows. Capturing these patterns requires architectures capable of global temporal reasoning without sequential bottlenecks. Second, visual and behavioral signals are not independent; they interact in non-linear and context-sensitive ways. A robust engagement estimation system must therefore learn cross-modal relationships rather than treating modalities as isolated feature streams [18]-[21].

Transformer-based architectures offer a principled solution to these challenges. Originally introduced in natural language processing, the Transformer replaces recurrence with self-attention mechanisms that directly model pairwise interactions across sequence elements. Self-attention enables global dependency modeling by computing relevance scores between all positions in a sequence simultaneously. This design eliminates sequential constraints, improves parallelization, and enhances long-range dependency capture. In computer vision, the vision transformer (ViT) extends this paradigm by dividing images into patches and processing them as token sequences, thereby enabling global spatial attention. Unlike CNNs, ViTs do not rely on locality-biased convolutional filters; instead, they learn contextual relationships across the entire image from the outset [22]-[25].

Motivated by these advantages, this paper proposes a cross-modal attention fusion framework built upon ViT for robust student attentiveness estimation. The proposed architecture replaces conventional convolutional and recurrent modules with transformer-based spatial and temporal modeling components. Visual frames are processed through a ViT backbone to capture global spatial dependencies. Behavioral features, including gaze direction, head pose, and blink dynamics, are embedded into a shared latent space. A cross-modal multi-head attention mechanism dynamically learns interactions between visual and behavioral modalities, enabling adaptive feature weighting based on contextual relevance. Temporal dependencies are subsequently modeled using a Transformer encoder, allowing comprehensive sequence-level reasoning without recurrence.

2. LITERATURE REVIEW

Automated student attentiveness detection has evolved significantly with the growth of online and hybrid learning environments. Early research primarily focused on visual observation and behavioral analysis within structured classroom settings. Negron and Graves [1] introduced the classroom attentiveness classification tool (ClassACT), a system designed to categorize attentiveness levels using visual monitoring techniques. Their framework demonstrated the feasibility of automated engagement classification; however,

it relied on handcrafted features and rule-based decision mechanisms, limiting adaptability in unconstrained online learning scenarios.

With the increasing adoption of e-learning platforms, research attention shifted toward behavioral monitoring in virtual environments. Revadekar *et al.* [2] proposed a system for gauging student attention using facial detection and activity tracking in online classrooms. Although their work emphasized continuous engagement monitoring, it lacked advanced temporal modeling strategies. Shah *et al.* [3] further extended behavioral analysis by incorporating gaze direction, head orientation, and facial expressions to assess attentiveness during e-learning sessions. While effective in demonstrating the importance of behavioral cues, the study employed conventional machine learning classifiers without modeling long-range temporal dependencies.

Subsequent studies emphasized face-based and ocular feature extraction for attentiveness detection. Pandey *et al.* [4] introduced a face detection-based attentiveness measure tailored for classroom environments. Their approach focused on identifying facial presence and orientation but remained limited to frame-level inference. Pai *et al.* [5] utilized CNNs for real-time eye monitoring to detect drowsiness and attentiveness. Although CNNs improved spatial feature learning, their locality-biased architecture restricted global contextual modeling.

Head pose estimation has also been explored as a strong indicator of engagement. Pinzon-Gonzalez and Barba-Guaman [6] employed head position estimation to determine attention levels in remote classrooms. Their findings validated head orientation as a reliable behavioral cue; however, the approach treated modalities independently and did not model inter-modal interactions. More comprehensive modeling strategies were later proposed by Elbawab and Henriques [7], who combined emotional and non-emotional features using machine learning techniques to enhance attentiveness prediction. While incorporating affective signals improved performance, the reliance on traditional learning pipelines limited the ability to capture complex temporal and cross-modal dependencies.

From an evaluation standpoint, reliability assessment remains essential in attentiveness monitoring systems. The intraclass correlation coefficient (ICC) framework proposed by Bi and Kuesten [8] provides a statistical foundation for assessing inter-rater consistency and model reliability. Such evaluation strategies are relevant when validating engagement detection models against annotated ground truth labels.

3. METHOD

The proposed framework is designed to estimate student attentiveness from video streams in both real-time and offline modes. The system processes input video frames to extract visual and behavioral cues, fuses these modalities using attention-driven mechanisms, models temporal dependencies through a transformer encoder, and outputs a probabilistic attentiveness score.

The overall pipeline consists of the following stages:

a) Frame acquisition and preprocessing
b) ViT-based spatial feature extraction
c) Behavioral feature embedding (gaze and head pose)
d) Cross-modal multi-head attention fusion
e) Temporal transformer encoding
f) Attentiveness classification

The dataset is the dataset for affective states in e-environments (DAiSEE), 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, very high). The DAiSEE dataset provides external validation and ensures the model generalizes across diverse learning contexts.

The proposed architecture follows a dual-stream transformer-based design that explicitly models complementary visual and behavioral cues before integrating them through attention-driven fusion. The framework is structured to capture spatial context, behavioral dynamics, and long-range temporal dependencies in a unified end-to-end trainable system.

3.1. Visual stream

The visual stream is responsible for extracting rich spatial representations from raw RGB frames captured during online learning sessions. Each input frame is first resized and normalized, then partitioned into fixed-size non-overlapping patches. If the frame resolution is H×W, and the patch size is P×P, the image is decomposed into $N = \frac{HW}{P^2}$ patches. Each patch is flattened and linearly projected into a high-dimensional embedding space through a learnable projection matrix, forming a sequence of patch tokens.

To retain spatial structure, positional encodings are added to each token embedding. These enriched tokens are then passed through a stack of ViT encoder blocks. Each block consists of multi-head self-attention (MHSA) and feed-forward layers with residual connections and layer normalization. The self-attention mechanism allows each patch token to attend to all other patches within the frame, thereby modeling global spatial dependencies. This is particularly important for attentiveness estimation, where facial cues, eye regions, and contextual posture information may interact non-locally.

The output of the ViT backbone is a refined visual token representation that captures holistic spatial context. A special classification token or pooled representation is extracted to summarize the frame-level visual information. This representation encodes facial orientation, eye region patterns, posture indicators, and other appearance-based cues relevant to attentiveness.

3.2. Behavioral stream

The behavioral stream operates in parallel to explicitly model structured behavioral features derived from the same video frame. Instead of relying solely on implicit learning from pixels, this stream incorporates interpretable behavioral indicators such as gaze direction vectors and head pose angles (yaw, pitch, and roll). The gaze vector, typically represented in 3D space, captures where the student is looking relative to the camera or screen. Head pose parameters quantify orientation deviations that may indicate distraction. These features are concatenated to form a behavioral feature vector at each time step.

To align the behavioral features with the visual embedding space, the concatenated vector is passed through linear projection layers. These fully connected layers transform low-dimensional behavioral inputs into a higher-dimensional behavioral embedding compatible with transformer attention operations. Non-linear activation functions and normalization may be applied to enhance representation capacity and stabilize training. The resulting behavioral embedding provides compact, structured information about student engagement that complements the high-dimensional visual token representation.

3.3. Cross-modal fusion and temporal modeling

After independent encoding, the visual token representation and behavioral embedding are fused using cross-modal multi-head attention. In this mechanism, one modality (e.g., visual tokens) serves as queries, while the other modality (behavioral embeddings) provides keys and values. This enables adaptive weighting of behavioral cues conditioned on visual context, allowing the model to dynamically determine the relative importance of gaze and head pose information for each frame. The fused features are then organized into temporal sequences and passed to a transformer-based temporal encoder. Unlike recurrent architectures, the temporal transformer models long-range dependencies using self-attention across time steps, enabling the system to capture sustained engagement patterns and gradual shifts in attentiveness. Finally, the temporally encoded representation is processed through a classification head with a Softmax layer to produce attentiveness probabilities. This hierarchical design ensures robust spatial modeling, explicit behavioral integration, and efficient sequence learning within a fully attention-driven framework.

3.4. ViT backbone

To overcome locality constraints of convolutional networks, the proposed framework employs a ViT backbone for spatial feature extraction. Each input frame $X \in R^{H \times W \times C}$ is divided into non-overlapping patches of size P×P. Each patch is flattened and linearly projected into a latent embedding space.

$$X_p = Flatten(Patch_i)W_e + E_{pos}$$

W_e is the learnable projection matrix
E_{pos} represents positional encoding
X_p denotes patch embeddings

This transforms the image into a sequence of tokens suitable for transformer processing. Each transformer encoder block computes self-attention across all patch tokens, enabling global spatial dependency modeling.

The attention mechanism is defined as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where:
Q, K, V are query, key, and value matrices
d_k is the key dimension
This allows each patch to attend to every other patch, capturing holistic contextual relationships.

3.5. Behavioral feature embedding

Behavioral cues provide complementary information to visual appearance features. The proposed framework embeds gaze vectors and head pose parameters into a unified latent space.

Let the behavioral feature vector at time t be:

$$B_t = [G_t, H_t]$$

where:
G_t= gaze direction vector
H_t = head pose angles (yaw, pitch, roll)

3.6. Gaze vector encoding

The gaze direction is represented as a 3D vector and projected into the embedding dimension:

$$G_t' = W_g G_t + b_g$$

3.7. Head pose embedding

Head pose angles are similarly projected:

$$H_t' = W_h H_t + b_h$$

the final behavioral embedding is:

$$B'_t = [G'_t, H'_t]$$

This embedding aligns dimensionality with visual tokens for fusion.

3.7. Cross-modal attention fusion

Instead of static concatenation or weighted summation, the proposed framework employs cross-modal multi-head attention to dynamically model interdependencies between visual and behavioral modalities.

Given visual tokens X_t and behavioral embedding B'_t, cross-attention is computed as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where queries are derived from visual features and keys/values from behavioral embeddings (or vice versa depending on configuration).

3.8. Multi-head attention formulation

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O$$

Each head learns distinct inter-modal relationships, enabling:
- Adaptive modality weighting
- Context-aware feature interaction
- Robustness to noise in individual modalities

The fused representation is denoted as:

$$Z_t$$

3.9. Temporal transformer encoder

Attentiveness is inherently sequential. Instead of recurrent modeling, the proposed framework uses a transformer encoder to capture long-range temporal dependencies.

Given the fused sequence:

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

Temporal encoding is performed as:

$$Z' = TransformerEncoder(Z)$$

This mechanism:
- Enables global temporal attention
- Eliminates recurrence constraints
- Improves parallelization
- Captures long-term engagement patterns

Positional encodings are added to preserve temporal ordering.

3.10. Attentiveness classification head

The temporally encoded representation Z' is passed through a fully connected classification head. For each time step:

$$y_t = Softmax(W_o Z'_t + b_o)$$

where:
W_o: is the output projection matrix
b_o: is bias
y_t: is the attentiveness probability distribution
The final attentiveness score can be computed as:
- Per-frame probability
- Sliding window average
- Sequence-level aggregation

Compared to conventional CNN–LSTM pipelines, the proposed transformer-based framework introduces several architectural advantages that significantly enhance attentiveness estimation performance. First, by employing a ViT backbone instead of convolutional layers, the model captures global spatial dependencies across the entire frame through self-attention, rather than relying on the limited receptive fields of convolutional kernels.

4. RESULTS AND PERFORMANCE EVALUATION

This section presents the quantitative and comparative evaluation of the proposed Transformer-based cross-modal attentiveness estimation framework. The model is evaluated using standard classification metrics including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Performance is compared against baseline deep learning models and selected methods from prior literature. The dataset was divided into training (70%), validation (15%), and testing (15%) subsets using stratified sampling. All experiments were conducted using identical preprocessing and evaluation protocols to ensure fairness. The dataset was divided into training (70%), validation (15%), and testing (15%) subsets using stratified sampling. All experiments were conducted using identical preprocessing and evaluation protocols to ensure fairness.

4.1. Overall classification performance

To evaluate the complete framework, all primary classification metrics on the test set are summarized in Table 1. The high recall indicates strong detection capability for attentive students, while high specificity confirms reliable identification of inattentive states. The balanced F1-score demonstrates stable performance across both classes.

Table 1. Overall performance metrics

Metric	Proposed framework
Accuracy	94.87%
Precision	93.95%
Recall (Sensitivity)	95.42%
Specificity	93.81%
F1-Score	94.68%
AUC	0.972

4.2. Ablation study

To quantify the impact of architectural components, a detailed ablation study was conducted as shown in Table 2. All evaluation metrics are reported for fairness. The results demonstrate that each architectural enhancement contributes consistently across all evaluation metrics.

Table 2. Ablation study with multiple metrics

Model variant	Accuracy	Precision	Recall	F1-Score	AUC
CNN + LSTM (Baseline)	86.32%	85.74%	86.10%	85.91%	0.884
Vision transformer only	90.14%	89.62%	90.31%	89.72%	0.921
ViT + Behavioral (Concatenation)	92.08%	91.73%	91.58%	91.65%	0.943
ViT + Cross-Modal Attention	93.11%	92.85%	92.58%	92.70%	0.956
Full proposed framework	94.87%	93.95%	95.42%	94.68%	0.972

4.3. Comparison with existing methods

To position the proposed framework against prior literature, multiple performance metrics are compared with representative approaches as in Table 3. The proposed method consistently outperforms existing techniques across all evaluation metrics, not only accuracy. The improvement in recall indicates better detection of engaged students, while higher precision reduces false engagement predictions.

Table 3. Comparative performance with existing studies

Method	Accuracy	Precision	Recall	F1-Score
Elbawab and Henriques [7]	89.30%	88.92%	88.60%	88.75%
Shah *et al.* [3]	85.60%	84.88%	85.10%	84.92%
Pandey *et al.* [4]	83.40%	82.95%	82.80%	82.88%
Proposed framework	94.87%	93.95%	95.42%	94.68%

4.4. Computational efficiency

To ensure practical feasibility, computational metrics are also analyzed as shown in Table 4. Despite a higher parameter count, the proposed framework maintains competitive inference speed due to parallel attention computation. The results confirm the effectiveness of the proposed Transformer-based cross-modal framework for student attentiveness estimation. The model achieves superior accuracy and F1-score compared to baseline and existing approaches, indicating strong overall classification capability and balanced error distribution. In addition to accuracy, the framework maintains stable precision and recall values, demonstrating its ability to correctly identify both attentive and inattentive students without bias toward a specific class. The high AUC further reflects strong class separability and reliable discrimination across decision thresholds. The inclusion of transformer-based temporal encoding significantly enhances performance by capturing long-range engagement patterns that static frame-level models cannot represent. Importantly, these performance gains do not compromise computational feasibility. Despite increased architectural complexity, the model preserves practical inference efficiency through parallel attention mechanisms, making it suitable for real-time or near real-time deployment in online learning environments.

Table 4. Model complexity and inference performance

Model	Parameters (M)	Inference time (ms/frame)	Throughput (FPS)
CNN + LSTM	12.4	18.7	53
Vision transformer	21.8	16.2	61
Proposed framework	24.6	17.4	57

5. CONCLUSION

This paper presented a Transformer-based cross-modal framework for robust student attentiveness estimation in online learning environments. The proposed architecture integrates a ViT backbone for global spatial dependency modeling, behavioral feature embedding for gaze and head pose representation, cross-modal multi-head attention for adaptive feature fusion, and a transformer-based temporal encoder for long-range sequence modeling. Unlike conventional CNN–LSTM architectures, the framework is fully attention-driven, enabling dynamic multimodal interaction without recurrent bottlenecks. Quantitative evaluation demonstrates the effectiveness of the proposed method. The model achieved 94.87% accuracy, 93.95% precision, 95.42% recall, 94.68% F1-score, and an AUC of 0.972 on the test set. Compared to the CNN–LSTM baseline (86.32% accuracy, 85.91% F1-score), the proposed approach improved performance by more than 8% in accuracy and nearly 9% in F1-score. The ablation study further confirmed incremental gains from cross-modal attention and temporal transformer encoding. Additionally, the model maintained practical inference efficiency (17.4 ms per frame), supporting near real-time deployment.

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	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓
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] T. P. Negron and C. A. Graves, "Classroom attentiveness classification tool (ClassACT): The system introduction," in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, Mar. 2017, pp. 26–29, doi: 10.1109/percomw.2017.7917513.

[2] A. Revadekar, S. Oak, A. Gadekar, and P. Bide, "Gauging attention of students in an e-learning environment," in *2020 IEEE 4th Conference on Information & Communication Technology (CICT)*, Dec. 2020, pp. 1–6, doi: 10.1109/cict51604.2020.9312048.

[3] 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)*, Jan. 2021, pp. 1–7, doi: 10.1109/iccci50826.2021.9402283.

[4] R. K. Pandey, A. A. Faridi, and G. Shrivastava, "Attentiveness measure in classroom environment using face detection," in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, Jan. 2021, pp. 1053–1058, doi: 10.1109/icict50816.2021.9358600.

[5] R. Pai, A. Dubey, and N. Mangaonkar, "Real time eye monitoring system using CNN for drowsiness and attentiveness system," in *2021 Asian Conference on Innovation in Technology (ASIANCON)*, Aug. 2021, pp. 1–4, doi: 10.1109/asiancon51346.2021.9544624.

[6] J. G. Pinzon-Gonzalez and L. Barba-Guaman, "Use of head position estimation for attention level detection in remote classrooms," in *Proceedings of the Future Technologies Conference (FTC) 2021, Volume 1*, Springer International Publishing, 2021, pp. 275–293.

[7] M. Elbawab and R. Henriques, "Machine learning applied to student attentiveness detection: using emotional and non-emotional measures," *Education and Information Technologies*, vol. 28, no. 12, pp. 15717–15737, May 2023, doi: 10.1007/s10639-023-11814-5.

[8] J. Bi and C. Kuesten, "Intraclass correlation coefficient (<scp>ICC</scp>): A framework for monitoring and assessing performance of trained sensory panels and panelists," *Journal of Sensory Studies*, vol. 27, no. 5, pp. 352–364, 2012, doi: 10.1111/j.1745-459x.2012.00399.x.

[9] S. T, A. B. J, V. V, B. A. J, G. V. A. S, and D. K, "Assessment of student attentiveness in classroom environment using deep learning," in *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)*, Aug. 2023, pp. 26–30, doi: 10.1109/iccpct58313.2023.10245194.

[10] S. Gupta and P. Kumar, "Attention recognition system in online learning platform using EEG signals," in *Emerging Technologies for Smart Cities*, Springer Singapore, 2021, pp. 139–152.

[11] 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, Jan. 2021, doi: 10.1007/s00521-020-05587-y.

[12] M. Sari, A. Moussaoui, and A. Hadid, "A simple yet effective convolutional neural network model to classify facial expressions," in *Modelling and Implementation of Complex Systems*, Springer International Publishing, 2020, pp. 188–202.

[13] A. A. Ravindran, "Internet-of-things edge computing systems for streaming video analytics: trails behind and the paths ahead," *IoT*, vol. 4, no. 4, pp. 486–513, Oct. 2023, doi: 10.3390/iot4040021.

[14] A. I. Wang and R. Tahir, "The effect of using Kahoot! for learning – A literature review," *Computers & Education*, vol. 149, p. 103818, May 2020, doi: 10.1016/j.compedu.2020.103818.

[15] 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, p. 5314, Aug. 2021, doi: 10.3390/s21165314.
[16] 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, Oct. 2021, doi: 10.1007/s00530-021-00854-x.
[17] M.-E. Otgonbold *et al.*, "SHEL5K: an extended dataset and benchmarking for safety helmet detection," *Sensors*, vol. 22, no. 6, p. 2315, Mar. 2022, doi: 10.3390/s22062315.
[18] 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*, Oct. 2019, pp. 9–14, doi: 10.1145/3349614.3356027.
[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, Feb. 2022, doi: 10.1177/07356331211069424.
[20] 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, 2021, doi: 10.1080/10447318.2021.1938389.
[21] 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)*, Jan. 2021, pp. 3336–3341, doi: 10.1109/icpr48806.2021.9411983.
[22] R. Periyasamy, M. Govindaraji, I. Nasurulla, V. Srinivasan, and K. Rama Devi, "Optimizing call center agent efficiency through deep learning-based classifications using SMFCCAE," *International Journal of Reconfigurable and Embedded Systems (IJRES)*, vol. 15, no. 1, 2026
[23] 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, p. 48, Mar. 2023, doi: 10.3390/bdcc7010048.
[24] Z. Wang, F. Zeng, S. Liu, and B. Zeng, "OAENet: oriented attention ensemble for accurate facial expression recognition," *Pattern Recognition*, vol. 112, p. 107694, Apr. 2021, doi: 10.1016/j.patcog.2020.107694.
[25] S. Hussain and J. Jeyachidra, "A stacked classifier model for enhanced student performance prediction in e-learning environments," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 14, no. 1, 2026.

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. Praveen Sundar MCA, M.Phil., Ph.D. is an Indian academician and researcher currently serving as Associate Professor & 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. Dr. Sundar 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. Dr. Sundar has also filed a patent related to an *"Intelligent Drug Abuse Ascertain System*. He can be contacted at email: praveensundarpv@gmail.com.