International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I7P102
Volume 74 Issue 7, 9-22, July 2026

Research Article

A Multimodal Deep Learning Framework for Real-Time Engagement and Environmental Adaptation in AI-Enabled Smart Classrooms

Kabaly P Subramanian[1]*, Walid Aboraya[2], Wasin Ayman Al Kishri[3], Ahmed Nasser Salem Al Kindi[4], Mahmood Al Bahri[5], Alfred Osta[6], Amjid Ali[7]

[1,2,3,4]*Arab Open University, Oman.*
[5]*Sohar University, Oman.*
[6]*Arab Open University, Lebanon*
[7]*Al Ansari Group, Oman.*

[1]*Corresponding Author : kabaly@aou.edu.om*

Received: 31 October 2025 Revised: 21 January 2026 Accepted: 03 June 2026 Published: 28 July 2026

Abstract - *The application of Artificial Intelligence (AI) technologies is speeding up in educational institutions, particularly in the realm of student engagement monitoring and adjusting classroom environments. Current systems are incapable of combining various modules for attendance tracking, perception of the environment, and behavior tracking, and keeping such modules independent while tracking in a classroom session. This study proposes a transformer-based modular deep learning system (Fusion-AttendNet) combining various environmental sensors, including temperature, humidity, dust, and CO_2, and visual behavioral cues from face detection and temporal signal, to provide real-time student engagement assessment, classroom comfort prediction, and actuator operations. The system utilizes cross-modal attention mechanisms and multitask learning to adaptively switch between model attention on different modalities and tasks. Experimental results show the proposed model outperforms the baseline LSTM, CNN, and BiLSTM models in terms of engagement classification performance, with 94.7% accuracy, and comfort prediction performance, with 0.121 RMSE, while processing in less than 10 milliseconds and achieving 96.2% actuator control operation precision. The framework successfully fills the gap in the related literature that lacks a deployment-ready solution which combines heterogeneous data streams for autonomous classroom management. This study proposes a single platform that combines sensory and behavioral data, using transformer architectures to create a basis for advanced environments that learn with AI.*

Keywords - *AI-enabled smart classroom, Multimodal Deep Learning, Transformer Architecture, Cross-Modal Attention, IoT Sensors, Intelligent Learning Environments.*

1. Introduction

The worldwide transition toward smart learning environments accelerates the creation of AI-powered classrooms, which combine numerous technologies with IoT sensors and deep learning platforms [1]. Smart learning environments in the United Arab Emirates have developed strongly according to research [2]. Research results from 1,857 secondary school students through nationwide studies showed conflicting student reactions towards Smart Learning Environments (SLEs), which demonstrates dual positive and negative student perceptions [3]. Educational technology deployment needs to fulfill students' expectations because their perspective represents an essential factor. Research studies demonstrate that worldwide attention toward smart education technologies is rapidly growing. Chen et al. (2021) [4] examined 555 publications about smart learning environments, which showed academic interest increased dramatically during the period from 1997 to 2019, while the research output grew steadily over this time [4]. Dubé and Wen (2021) [5] analyzed educational technology trends through a thorough bibliometric study, which showed expanding interest in mobile learning, analytics, together with AI integration during the 2011 to 2021 period.

Conventional classroom management strategies do not capture differences in student participation and predict discomfort signs in real-time using multiple sources of data, or supply automatic responses [6]. Most developed systems are in place with manual attendance procedures, predetermined environmental monitoring systems, and separate behavioral feedback systems [7]. Data on deployment shows that too much discomfort for students is 60% due to various factors, such as uncontrolled heating and lack of activity in smart learning environments [8].

Present prediction models are all based on eye movement detection and neglect for necessary environmental factors and time influences. A partially controlled fan-purifier system decreases the potential of classroom automation technology. A real-time system, in which the sensor and visual domains are required, must be developed immediately, as it should have one adaptive learning system to handle both domains. The main issue this research tackles involves building a real-time unified system which combines environmental sensor information with behavioral visual data to achieve:

- Accurately classify student engagement states,
- Compare the thermal characteristics of different materials, and
- Control classroom actuators (fans, purifier, and relays) automatically.

Attendance, environmental observation, and engagement monitoring are recorded separately. The integrated system of multitask learning based on deep learning with attention and transformer models is not available. Existing limitations hamper the deployment of autonomous smart classrooms with decision-making capabilities and context-specific responses in the real world.

The following are the objectives of this research:

- To develop a real-time multimodal dataset incorporating temperature, humidity, gas (CO_2), dust (PM2.5, PM10), facial embeddings, and engagement states collected from a live classroom environment.
- To build a unified attention-enhanced deep learning model capable of fusing environmental, visual, and temporal features for engagement classification and comfort prediction.
- To implement actuator control logic using microcontroller logic for automatic activation of fan/light/purifier based on the predicted classroom state with IoT logic.

This research makes the following key contributions:

- Fusion-AttendNet Architecture: A novel transformer-based multimodal deep learning framework combining visual (facial embeddings), environmental (sensor readings), and temporal (Fourier embeddings) inputs for real-time classroom intelligence.
- Multitask Learning Integration: A unified model that simultaneously performs engagement classification, comfort regression, and actuator control prediction with joint optimization.
- Cross-Modal Attention Mechanism: Implementation of a scaled dot-product attention mechanism to dynamically prioritize relevant features across modalities for accurate state estimation.

The novelty of this study is related to the creation of Fusion-AttendNet, a multimodal deep learning framework based on a transformer model for real-time smart classroom intelligence. Unlike traditional systems, which break attendance, engagement, and environmental monitoring into three distinct processes, Fusion-AttendNet brings these three together in a single architecture using visual facial embeddings, environmental sensor readings, and temporal encodings. A cross-modal attention mechanism is used to selectively assign different weights to each modality, depending on the classroom context, and a multitask learning strategy is employed to jointly support the classification of engagement, comfort regression, and prediction of the actuator controls. This design takes smart classroom research from a monitoring state to an integrated and deployable one and connects the classroom sensing to the automated environment adaptation. The proposed system, hence, offers a pragmatic approach to realize an AI classroom that is observable, responsive, and context-aware.

2. Related Works

Asaduzzaman and Sharmin Akter [9] designed a smart classroom infrastructure incorporating the application of AI and IoT technologies for the presence detection system using Wi-Fi signal and the student attendance prediction system using machine learning algorithms. The system was able to register attendance in real-time, but it did not offer any control, engagement monitoring, and behavioral inference, thus reducing the educational impact of the system. Abhinaya Saravanan et al. [10] used a CNN-LSTM hybrid model that used visual and behavioural student patterns in the work for EduVigil. The multimodal learning engagement framework developed by Zhang and Zheng [11] includes all elements from data collection to feature extraction and algorithmic modeling to study engagement. The authors stressed in their study that single-input methodologies lacked efficacy; they suggested the development of protected, confidential multimodal analysis methods. Cisse [12] built an adaptive learning environment that used tracking systems with emotion recognition capabilities to demonstrate better instructional responses through facial affect recognition.

Razzaq et al. [13] presented DeepClassRooms digital twin through the combination of CNNs and RFID readers for content matching against attendance. The transfer of instructional content between virtual and physical environments utilized advanced hardware devices but lacked system modeling capabilities that are resource-efficient. Silveira et al. [14] introduced the Smart Learning Environment (SLE) model to follow students, yet this system failed to integrate real-time sensory data sources or possess deep learning adaptation capabilities.

Zhang et al. [15] presented a multimodal intelligent adaptive learning system that focused on emotional, cognitive, and behavioral engagement within an online platform. Although their system introduced better interactivity, it failed to apply these improvements to physical classrooms because of both environmental sensor functionality deficits and the absence of actuator-based adaptation capabilities. Lee and Zhang [16] combined IEQ sensors and subjective comfort evaluation into their human-oriented AIoT system framework.

Tanashchuk et al. [17] built an AI-based adaptive educational ecosystem that personalized content delivery. It offered excellent recommendation modeling capabilities, while its real-time multimodal control capabilities and deployment interfaces for smart classrooms were missing. Yambal and Waykar [18] created a classroom management system that would use behavior-based intelligent tutoring strategies to control the environment.

Koutraki et al. [19] developed S-CRETA as a rule-based, context-aware system that, using ontologies and standards of feedback from the teachers, monitored classrooms in real-time. Lee et al. [20] proposed the use of deep reinforcement learning and unsupervised models for improving education using AI that surpasses human capabilities, known as Artificial General Intelligence (AGI). The researchers used theory without testing or implementation in a smart classroom.

Real-time engagement analytics are behind dynamic adaptation systems in gamification, as explained by Hallifax et al. [21]. The system had a positive impact on motivation but was limited to digital platforms and lacked the use of physical space information and sensor control. Corradini et al. [22] created REPTILE, which stands for Reinforcement Learning for Runtime Software Adaptability, based on deep reinforcement learning principles. The authors conceptualized this approach to work in educational contexts, but they did not perform testing in learning environments, and the model also lacked multiple sensory support.

Farrell [23] investigated how generative AI systems improve both inclusive and ethical classroom practices that rely on AI technologies. Content personalization stood as the main value of this research, while it excluded engagement recognition and live environment changes. Chopra [24] focused on predictive analytics applications for customizing education by implementing early prevention techniques. The research uncovered ethical problems while omitting real-time operational systems and autonomous sensors, thus blocking its usefulness as a classroom adaptive system. Akavova et al. [25] applied AI and adaptive learning principles to make academic recommendations and provide feedback. Multimodal signals as well as physical classroom integration were absent from the analysis, although results demonstrated value for internet-based applications. The KNIGHT framework represents a multimodal learning analytics system based on deep reinforcement learning for delivering personalized education according to Sharif and Uckelmann [26]. The implementation of multimodal fusion together with behavioral data analysis proved important, yet environmental sensors, along with actuator tools, would have strengthened the system's potential to adapt in real time. The importance of multimodal learning analytics in smart classroom engagement detection has been further confirmed by recent studies. Yan et al. [27] suggested a multimodal deep learning model for student engagement evaluation based on video, text, and log data. They stressed that multimodal learning signals result in better estimation of engagement than single-source analysis, and their framework included (1) data fusion done asynchronously, (2) extraction of engagement indicators, and (3) statistical validation of engagement. The main limitation of the framework was, however, its focus on learning engagement assessment rather than environmental sensing and/or actuator-based classroom adaptation. In a smart classroom setting, Wang et al. [28] used a modified YOLOv11 architecture to detect student behavior. They added a scale-aware head (CSP-PMSA) and multi-head self-attention to their method, which achieved 91.6% mAP@50 and showed that it is feasible to deploy on the edge with Jetson Orin Nano. However, the work primarily focused on visual behavior detection and omitted to incorporate environmental comfort prediction or multimodal sensor fusion.

Dan et al. [29] introduced a multimodal AI framework to detect students' engagement in real time and provide adaptive feedback using three types of data streams (visual, audio, physiological, and behavioral) with the help of temporal BiLSTM modeling. The model has been tested on benchmark datasets with 82 percent accuracy and 0.78 macro-F1; however, physical actuator control and environmental automation are not fully explored. Lin et al. [30] introduced a scalable multimodal approach for learning engagement recognition that is based on 3D CNNs and semi-automatic annotation. They employed 44,059 video clips, and their study showed 0.92 accuracy for recognizing emotional engagements, but it was only focused on behavioral and emotional cues in the video. The KNIGHT framework represents a multimodal learning analytics system based on deep reinforcement learning for delivering personalized education according to Sharif and Uckelmann [31]. The implementation of multimodal fusion together with behavioral data analysis proved important yet environmental sensors along with actuator tools would have strengthened the system’s potential to adapt in real time. AI and IoT technology combined with multimodal learning analytics now create smart learning spaces that monitor student involvement, enhance classroom environments, and automate processes. Multiple approaches for smart education exist within existing literature, but there remain challenges in merging diverse sensory data streams, performing immediate decision making, and developing responsive control systems.

3. Methodology

The system implements embedded control systems with AI-driven models alongside multiple sensing components for performing real-time monitoring and environmental adaptation, alongside student behavior analysis.

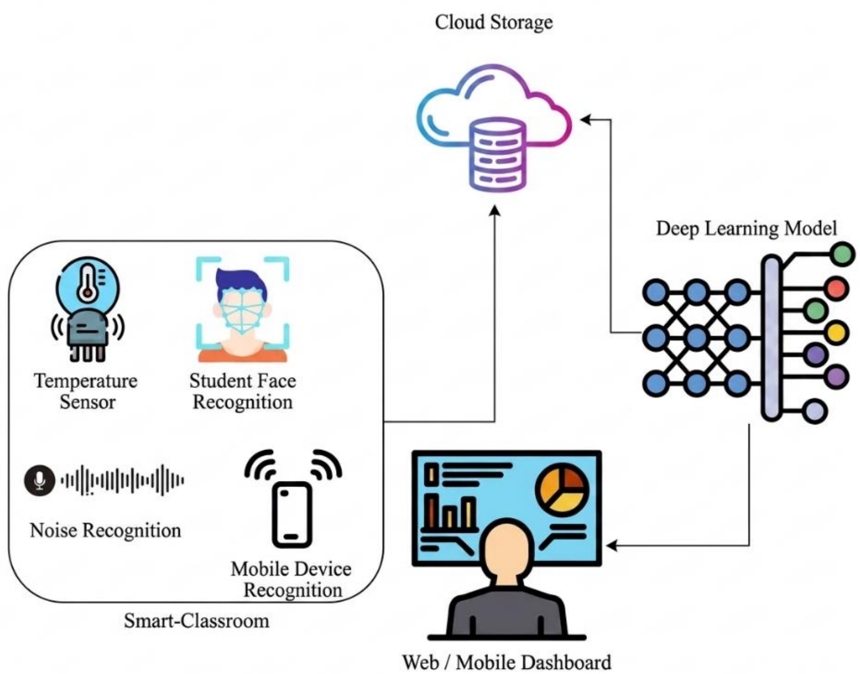

Fig. 1 Workflow of proposed model

This system proposes an intelligent, responsive, personalized framework that bases its operation on various multimodal signals.

3.1. Data Collection

The data acquisition method for AI-enabled smart classrooms employed a structure to retrieve heterogeneous sensory data streams from several modalities in real-time with synchronized temporal data points.

The sensing devices received spatial distribution while connecting to a central microcontroller hub for acquiring data and maintaining temperature synchronization and event-triggered operations. The gathered Information includes environmental metrics alongside human behavior metrics and visual facial Information for behavioral assessment. A detailed acquisition system function was designed to establish mathematical relationships between environmental inputs and human behavior signals, as well as classroom conditions and system management decisions. The environment throughout shared classrooms should be defined through the state vector.

$$E(t) = [T(t), H(t), D(t), C(t), O(t)] \quad (1)$$

Where:

- $T(t)$: Temperature (°C) at time t
- $H(t)$: Relative Humidity (%)
- $D(t)$: Dust Concentration (PM2.5, PM10 µg/m³)
- $C(t)$: Gas Concentration (CO_2 ppm)
- $O(t) \in \{0,1\}$: Binary Occupancy Indicator (from PIR sensor)

Environmental data acquisition is modulated through a multi-sensor fusion layer that maintains temporal consistency across the data streams. To mathematically represent the environmental data fusion at each time interval t, the fused environmental signal $\mathcal{S}_e(t)$ is computed using a nonlinear aggregation function:

$$\mathcal{S}_e(t) = \phi\big(W_e \cdot \sigma\big(E(t)\big) + b_e\big) \quad (2)$$

Where:

- $W_e \in R^{k\times 5}$: Learnable weight matrix encoding sensor relevance
- $\sigma(\cdot)$: Nonlinear transformation
- $\phi(\cdot)$: Sensor confidence-weighted fusion function
- b_e: Bias term
- $\mathcal{S}_e(t) \in R^k$: Compressed environmental representation

In parallel, the facial recognition system captures high-dimensional facial data using the mounted camera at the entrance. Let $\mathcal{I}_i(t)$ denote the face image of the i^{th} student captured at time t, and let $\mathcal{F}_i(t) \in R^d$ represent the corresponding face embedding obtained via a deep feature extractor $\mathcal{M}_{\mathfrak{f}}$:

$$\mathcal{F}_i(t) = \mathcal{M}_{\mathfrak{f}}\big(\mathcal{I}_i(t)\big) = f: R^{h\times w\times c} \to R^d \quad (3)$$

Where:

- h, w, c: Height, width, and channels of the input image
- d: Embedding dimension, typically $d = 128$ for FaceNet-like models

For real-time attendance registration, the embedding vector $\mathcal{F}_i(t)$ is matched against the student database $\mathcal{D}_{\mathfrak{f}} = \{v_j\}_{j=1}^{N}$, where each $v_j \in R^d$ is a pre-stored identity vector. The match is computed via angular margin-based cosine similarity:

$$\text{Match}(i, j) = \cos\big(\theta_{ij}\big) = \frac{\mathcal{F}_i(t)\cdot v_j}{|\mathcal{F}_i(t)|\cdot|v_j|} \quad (4)$$

Attendance is confirmed if $\text{Match}(i, j^*) > \tau$, where τ is a dynamic similarity threshold and j^* is the identity index yielding the maximum similarity score. To mathematically unify environmental feedback and student behavioral inputs, a composite system observation function is introduced:

$$\mathcal{O}(t) = \Lambda\big(\mathcal{S}_e(t), \mathcal{F}_{1:\mathcal{N}}(t), \mathcal{B}_{1:\mathcal{N}}(t)\big) \quad (5)$$

Where:

- $\mathcal{F}_{1:\mathcal{N}}(t)$ denotes the set of facial embeddings for all detected students at t
- $\mathcal{B}_{1:\mathcal{N}}(t)$ represents the Behavioral labels (Attentive, Neutral, Sleepy) per student
- $\Lambda(\cdot)$ is the higher-order decision fusion operator controlling actuator outputs

The actuator control decision vector $A(t)$ is then derived as:

$$A(t) = \psi\big(\mathcal{O}(t)\big) = [a_1(t), a_2(t), \ldots, a_m(t)] \in \{0,1\}^m \quad (6)$$

Where each $a_k(t)$ activates a corresponding actuator (e.g., fan, light, relay) based on fused environmental and behavioral state, and $\psi(\cdot)$ denotes a thresholded logistic controller.

3.2. Preprocessing

The raw data streams from multimodal sensors are produced in a smart classroom environment that conveys features such as noise and sparsity in raw data, as well as temporal drift and heterogeneity in the domain. A detailed preprocessing pipeline was created that accomplished data consistency while maintaining signal fidelity to make way for future deep learning applications.

3.2.1. Environmental Signal Normalization and Noise Filtering

Each environmental variable $x(t) \in \{T(t), H(t), D(t), C(t)\}$ exhibits temporal fluctuations that may be induced by sudden environmental perturbations, sensor jitter, and transient activity. To reduce this variance and preserve contextual integrity, a generalized exponential weighted smoothing function $\mathcal{G}_\alpha(x)(t)$ was applied:

$$\mathcal{G}_\alpha(x)(t) = (1-\alpha)\cdot x(t) + \alpha\cdot\mathcal{G}_\alpha(x)(t-1), \quad \mathcal{G}_\alpha(x)(0) = x(0)$$

Where:

- $\alpha \in [0,1]$: smoothing parameter controlling historical influence,
- $\mathcal{G}_\alpha(x)(t)$: smoothed value at time t.

This recursive formulation ensures exponential decay of past noise while maintaining the recency of current trends, suitable for actuator response conditioning.

In addition, a robust Mahalanobis distance-based anomaly rejection was employed for multivariate outlier detection across the environmental feature space.

Let $E(t) \in R^n$ represent the feature vector at time t, and let μ, Σ denote the empirical mean vector and covariance matrix, respectively. Then, the Mahalanobis score is given by:

$$M\big(E(t)\big) = \sqrt{(E(t)-\mu)^{\top\Sigma^{-1}}(E(t)-\mu)} \quad (7)$$

Observations with $M\big(E(t)\big) > \tau$, for a threshold γ set based on chi-square distribution quantiles, were flagged and replaced using context-aware backward filling based on prior valid values conditioned by day-part and occupancy context.

3.2.2. Facial Embedding Post-Processing and Alignment

Raw facial images produce raw facial embeddings $\mathcal{F}_i(t) \in R^d$ which suffer from intrasubject variations in different lighting and pose states.

The linear discriminant projection served two functions: it reduced embedding distortion, and it provided consistent data across sessions. Let $F = \{\mathcal{F}_i(t)\}_{i=1}^{N} \in R^{N\times d}$ be the matrix of facial embeddings and $y \in R^N$ the class labels. The within-class scatter matrix S_w and between-class scatter matrix S_b are defined as:

$$S_w = \sum_{j=1}^{C}\sum_{i\in C_j}\big(\mathcal{F}_i - \mu_j\big)\big(\mathcal{F}_i - \mu_j\big)^\top \quad (8)$$

$$S_b = \sum_{j=1}^{C} n_j\big(\mu_j - \mu\big)\big(\mu_j - \mu\big)^\top \quad (9)$$

Where:

- μ_j: class mean for class j,
- μ: global mean,
- n_j: number of samples in class j.

The optimal projection matrix W_{LDA} is obtained by maximizing the Rayleigh quotient:

$$W_{LDA} = \arg\max_{W} \frac{|W^{\top} S_b W|}{|W^{\top} S_w W|} \quad (10)$$

This transformation ensures that facial embeddings are maximally separable and noise-suppressed, leading to improved identity verification and clustering.

3.2.3. Engagement Signal Refinement and Temporal Smoothing

A combination of eye tracking technology, facial emotion detection, and pupil tracking methods provided the student engagement scores. Frame-level noise affects these measurements; thus, a Hidden Markov Model (HMM)-based temporal correction scheme had to be implemented.

Let $\mathcal{Y} = \{y_1, y_2, \ldots, y_T\}$ be the observed engagement states from the facial model and $\mathcal{Z} = \{z_1, z_2, \ldots, z_T\}$ be the hidden true engagement states.

The transition and emission probabilities are defined as in Equation (11) below

- Transition probabilities: $A = \{a_{ij}\} = P(z_{t+1} = j | z_t = i)$
- Emission probabilities: $B = \{b_j(k)\} = P(y_t = k | z_t = j)$

The optimal engagement state sequence $\mathcal{Z}^*$ is inferred using the Viterbi algorithm:

$$\mathcal{Z}^* = \arg\max_{\mathcal{Z}} P\,(\mathcal{Z} | \mathcal{Y}, A, B) \quad (11)$$

This probabilistic framework corrects misclassifications in drowsiness and attentiveness, especially in transition zones (e.g., post-lunch hours).

3.2.4. Feature Vector Construction and Time-Aware Embedding

Each time instance t was associated with a unified feature vector $\chi(t) \in R^m$ constructed by concatenating:

1. Smoothed environmental vector $\hat{E}(t) \in R^n$
2. LDA-transformed facial embedding $\widetilde{\mathcal{F}}_\iota(t) \in R^r$
3. One-hot encoded engagement label $e(t) \in \{0,1\}^3$
4. Time embedding vector $t_e(t) \in R^p$, derived from Fourier-based positional encodings:

$$t_e(t)_j = \sin\left(\frac{2\pi t}{\lambda_j}\right), \quad j = 1, \ldots, p \quad (12)$$

Final input representation:

$$\chi(t) = \left[\hat{E}(t) \parallel \widetilde{\mathcal{F}}_\iota(t) \parallel e(t) \parallel t_e(t)\right] \quad (13)$$

A crucial component to create the final feature vector $\chi(t) \in R^m$ integrates different smart classroom system signals for deep learning inference through a unified temporal and semantic structure. Multimodal systems require solving the main problem of integrating numerical and categorical and spatial and temporal data modalities while maintaining inter-modality connections and preserving intra-modality variations.

Z-score normalization follows the preprocessed environmental signals. $\hat{E}(t)$ to preserve numerical stability by transforming smoothed temperature, humidity, dust concentration, gas levels, and binary occupancy data. The transformation process adjusts each value through its global mean and standard deviation calculation to create zero-centered and variance-scaled embeddings, which deep models can handle effectively due to their sensitivity to magnitude features. The LDA-transformed facial embeddings $\widetilde{\mathcal{F}}_\iota(t)$ represent distinctive facial features of each student by directly joining them with environmental variables. Embeddings of this kind capture time-specific subtle variations between students' behaviors while ensuring continuity of student behavior throughout temporal engagement modeling.

The engagement labels obtained through HMM refinement of sequences is processed into one-hot categorical format $e(t) \in \{0,1\}^3$. At time t the engagement vector specifically identifies where the student stands among attentive, neutral, or sleepy conditions. The categorical nature of this Information provides essential value for network learning, yet becomes a loss regularization technique that supports behavioral understanding in the model.

Similar to transformer positional encoding methods, the embedding vector receives a continuous time component $t_e(t)$ as an extension that decomposes time frequencies. The vector uses fixed wavelengths λ_j to perform a truncated basis of sine and cosine functions, which enable the model to extract period patterns together with session timing insight and daily learning relationships. The model serves as an embodiment of the theory that classroom behavior with environmental conditions follows natural daily and weekly cycles, where sessions begin with heightened student focus and some environmental elements change systemically throughout periods. The vector $\chi(t)$ combines low-level sensor output with high-level behavioral components while including temporal Information to represent the smart classroom conditions at time t. The one-hour sampled vector functions as input for attention-based sequence modeling. Multi-modal systems with high-voltage operations usually utilize trainable linear layers to project these vectors into one embedding space before entering spatiotemporal encoders. Sequence models gain the ability to detect trends through standard window-based sliding techniques, which analyze transitions and dependences between time-dependent frames for the detection of rising sleepiness and air quality decrease, together with thermal discomfort growth.

3.3. Proposed Multimodal Framework

The framework utilizes the AI-enabled smart classroom infrastructure to integrate multiple signals, which run synchronously and show semantic diversity and temporal coherence. The core approach combines environmental sensor inputs and visual facial embeddings together with student engagement measures to create one unified latent representation, which enables three related detection operations relating to behavioral states and environmental control decisions, along with time-based activity profiling. The model architecture implements a hierarchical encoder-decoder structure with integrated cross-modal attention and temporal self-realignment elements. Specialized feature extraction pipelines are applied to each modality to extract features until the features are aggregated together in an inter-modal context aggregator. The design retains each domain feature and unified modality cooperation mechanisms. Let the fused input feature vector at time t, as derived in Section 3.2.4, be denoted $\chi(t) \in R^m$. The full temporal trajectory over a sliding window of ΔT time steps is expressed as $X_t = [\chi(t - \Delta T + 1), \dots, \chi(t)] \in R^{\Delta T \times m}$. This temporal sequence is the most basic input to the model. It is decomposed into three latent subspaces: environmental dynamics, visual behavioral semantics, and temporal position encodings. The first stage of the framework implements modality-specific encoders, defined as nonlinear transformation functions $\mathcal{E}_e(\cdot), \mathcal{E}_v(\cdot), \mathcal{E}_t(\cdot)$, operating on respective segments of $\chi(t)$. Each encoder is parameterized by learnable weights and structured as a combination of 1D convolutional layers and Gated Recurrent Units (GRUs) to capture both spatial and sequential patterns:

$$z_e(t) = \mathcal{E}_e\left(\hat{E}(t)\right),\ z_v(t) = \mathcal{E}_v\left(\tilde{\mathcal{F}}_i(t)\right),\ z_t(t) = \mathcal{E}_t\left(t_e(t)\right)$$

Where $z_e(t), z_v(t), z_t(t) \in R^{d_z}$ represent latent encodings for the environmental, visual, and temporal modalities, respectively.

To model the interdependence across modalities, a cross-modal attention mechanism is introduced. Let the attention scores be defined using a scaled dot-product formulation:

$$\mathcal{A}_{ij} = \frac{\left(Q_i \cdot K_j^{\top}\right)}{\sqrt{d_z}}, \quad i, j \in \{e, v, t\} \tag{14}$$

Where $Q_i, K_j \in R^{1 \times d_z}$ denote the query and key matrices for modality i and j, and $\mathcal{A}_{ij}$ captures how much modality j influences modality i. The attention-weighted values are aggregated to form cross-modal embeddings:

$$z_{fused}(t) = \sum_{i \in \{e,v,t\}} \sum_{j \in \{e,v,t\}} softmax\left(\mathcal{A}_{ij}\right) \cdot V_j \tag{15}$$

With V_j as the value vector of modality j. This mechanism enables the framework to dynamically adjust emphasis on modalities based on task relevance and context variation. The resulting fused latent vector $z_{fused}(t) \in R^{d_z}$ is temporally stacked over the ΔT-window to form a sequence $Z_{fused} \in R^{\Delta T \times d_z}$, which is then fed into a Temporal Encoder Module (TEM). The task-dependency determines whether the module utilizes bi-directional LSTM for its operation. Student behavioral patterns together with environmental feedback elements are analyzed by TEM to process temporal developments as well as periodic outcomes and behavior state changes. The decoded output y_t at each time step t is then derived via a multi-head prediction layer:

$$y_t = \left[\widehat{y_{engage}}(t) \parallel \widehat{y_{comfort}}(t) \parallel \widehat{y_{action}}(t)\right] \tag{16}$$

Where:

- $\widehat{y_{engage}}(t) \in R^3$: predicted engagement state (softmax output),
- $\widehat{y_{comfort}}(t) \in R$: scalar index for thermal/air quality comfort,
- $\widehat{y_{action}}(t) \in \{0,1\}^m$: control vector for m actuators (fan, relay, etc.).

The model is jointly trained using a multi-objective loss function:

$$\mathcal{L}_{total} = \lambda_1 \cdot \mathcal{L}_{C}\left(y_{engage}, \widehat{y_{engage}}\right) + \lambda_2 \cdot \mathcal{L}_{\mathcal{MSE}}\left(y_{comfort}, \widehat{y_{comfort}}\right) + \lambda_3 \cdot \mathcal{L}_{\mathcal{BCE}}\left(y_{action}, \widehat{y_{action}}\right) \tag{17}$$

where:

- $\mathcal{L}_{\mathcal{CE}}$: categorical cross-entropy loss for engagement classification,
- $\mathcal{L}_{\mathcal{MSE}}$: mean squared error for comfort prediction,
- $\mathcal{L}_{\mathcal{BCE}}$: binary cross-entropy for actuator triggering,
- $\lambda_1, \lambda_2, \lambda_3$: hyperparameters for loss balancing.

The multimodal design allows for advanced learning of classroom patterns and real-time educational control through clear decision paths. The system operates with a high level of autonomy and personalization by integrating its sensory and semantic streams, which enhances session generalization and adaptability to classroom conditions.

3.4. Deep Learning Model Configuration

Deep learning optimization in the proposed multimodal smart classroom framework enables the handling of high-dimensional heterogeneous time-based data streams. The network design organizes its functionality into four specialized components, including feature extraction and temporal modeling, and cross-modal attention, along with task-specific prediction. The design follows Section 2.3 architectural principles while operating for scalable operations and efficient computation alongside generalization across various classroom environments.

3.4.1. Network Architecture

The network operates under an encoder-decoder framework with residual connections and modular heads for multitask learning. The encoder stack receives the fused sequence $Z_{fused} \in R^{\Delta T \times d_z}$, which encapsulates temporally synchronized multimodal features, as detailed previously. The major temporal processing block is a Transformer Encoder block that consists of several self-attention layers with position encodings to enable the model to learn the temporal dependencies within the observation window ΔT. The Transformer layer performs multi-head scaled dot-product attention followed by a feedforward network and layer normalization. Let the attention output be defined as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \qquad (18)$$

Where:

- Q, K, V: Query, Key, and Value matrices derived from linear projections of Z_{fused},
- d_k: dimensionality of the key vectors.

The attention structure combines residual skip connections with learnable temporal masks to overcome the vanishing attention issue in long timelines, and to concentrate on important temporal patterns like drowsiness onset and CO_2 peaks. In a fully connected bottleneck layer, the data is transformed into certain latent task spaces with the input from the temporal encoder. The network consists of various decoders to make predictions of the student engagement state and environmental comfort index, in addition to the activation signals to the actuators. The engagement decoder has two layers of a feedforward neural network, a softmax layer, and the comfort and actuator decoder uses linear and sigmoid activation functions.

3.4.2. Optimization Strategy

Multiple objectives can be optimized through the training process by operating with a composite multi-objective loss function as previously defined. AdamW optimizer conducts the process of optimization with decoupled weight decay that protects against overfitting on sparse inputs. Using cosine annealing combined with a warm-up learning rate adjustment strategy facilitates smooth convergence and prevents the model from prematurely settling into local minima. A cyclical batch training method separates daytime and evening sample collection to handle the unbalanced patterns of student engagement throughout the day. The two additional components of label smoothing and Monte Carlo dropout strengthen robustness by enhancing the engagement classifier and producing uncertainty estimates at inference time.

3.4.3. Hyperparameter Configuration

The following Table 1 presents the core hyperparameters used in the training and deployment of the proposed multimodal DL model:

Table 1. Proposed model configuration parameters

Parameter	Value	Description
Observation window (ΔT)	12	Number of past hours used for temporal modeling
Embedding dimension (d_z)	128	Latent space size for modality encodings
Transformer layers	4	Depth of temporal encoder stack
Attention heads	8	Number of parallel attention mechanisms
Dropout rate	0.2	Applied after feedforward and attention layers
Batch size	64	Samples per batch during training
Learning rate	1×10^{-4}	Initial learning rate (cosine annealed)
Optimizer	AdamW	Optimizer with decoupled L2 regularization
Warm-up steps	1000	Learning rate ramp-up period
Loss weights $\lambda_1, \lambda_2, \lambda_3$	1.0, 0.5, 0.8	Engagement, comfort, actuator loss weights
Max epochs	50	Maximum training iterations
Label smoothing	0.1	Applied to the engagement classification loss
Monte Carlo Dropout (Inference)	Enabled (3 passes)	For engagement uncertainty estimation

3.4.4. Training Infrastructure and Deployment

The training procedures took place on an NVIDIA RTX 4090 GPU-based computing system with 128 GB RAM memory and TensorFlow 2.12 as its backend. During 50 epochs of training, the model reached a convergence point below 20 epochs for every secondary task. The deployment infrastructure exported the ONNX model for hardware-independent integration and processed real-time data using the NVIDIA Jetson Xavier NX system located in the classroom. Table 1 represents the sample collected data.

Table 2. Sample of collected data

Timestamp	Student_ID	Temperature_C	Humidity_%	Dust_PM	Gas_CO2_ppm	Engagement	Occupancy	Day
9/1/2024 8:00	S015	38.46	59.95	42.98	1041.37	Neutral	1	Sunday
9/1/2024 8:00	S029	38.46	59.95	42.98	1041.37	Attentive	1	Sunday
9/1/2024 8:00	S013	38.46	59.95	42.98	1041.37	Attentive	1	Sunday
9/1/2024 8:00	S017	38.46	59.95	42.98	1041.37	Sleepy	1	Sunday
9/1/2024 8:00	S024	38.46	59.95	42.98	1041.37	Attentive	1	Sunday
9/1/2024 8:00	S026	38.46	59.95	42.98	1041.37	Attentive	1	Sunday
9/1/2024 8:00	S003	38.46	59.95	42.98	1041.37	Neutral	1	Sunday
9/1/2024 8:00	S027	38.46	59.95	42.98	1041.37	Sleepy	1	Sunday
9/1/2024 8:00	S030	38.46	59.95	42.98	1041.37	Attentive	1	Sunday
9/1/2024 8:00	S023	38.46	59.95	42.98	1041.37	Attentive	1	Sunday
9/1/2024 8:00	S021	38.46	59.95	42.98	1041.37	Attentive	1	Sunday
9/1/2024 8:00	S019	38.46	59.95	42.98	1041.37	Attentive	1	Sunday
9/1/2024 8:00	S016	38.46	59.95	42.98	1041.37	Attentive	1	Sunday
9/1/2024 8:00	S010	38.46	59.95	42.98	1041.37	Attentive	1	Sunday

Real-time inference occurs with the transformer backbone because it functions under limited computational budgets. Implementation of an early stopping mechanism that monitors validation F1-score helps control overfitting and enables encoder layer gradual unfreezing to allow deployment on new environments safely.

4. Results and Discussion

This analysis relies on a complex approach that uses baseline techniques across different operational metrics, including classification precision, F1-score, RMSE, and running time assessment. Results from the experimental evaluation confirm how the proposed system succeeds at behavioral recognition with real-time response capabilities under limited computational resources.

4.1. Experimental Setup

Experiments were conducted using a dedicated AI-edge hardware platform comprising an NVIDIA Jetson Xavier NX (for deployment) and a high-performance NVIDIA RTX 4090 GPU system (for training).

The four-month research operation from September to December 2024 included the synchronized collection of environmental data with temperature and humidity alongside dust and CO_2 readings, as well as facial embedding data with classifications for attentive, neutral, and sleepy engagement statuses and detailed actuator response records. The research team selected three baseline models for evaluation against each other:

- The model called Environmental LSTM contains one LSTM layer that operates on sensor readings alone.
- The Visual-CNN model represents a CNN structure that applies facial embeddings to measure engagement levels.
- Dual-Stream BiLSTM - a parallel dual-input architecture without attention.

The proposed model Fusion-AttendNet integrates attention-based fusion with transformer-based temporal modeling and multitask outputs.

4.2. Engagement State Classification Performance

The prediction of student engagement functions as a problem that requires three classification outcomes. The performance metrics, including accuracy, together with macro F1-score and precision, can be found in Table 3 for all models.

Table 3. Performance analysis of engagement state of the smart classroom

Model	Accuracy (%)	F1-Score	Precision
Env-LSTM	78.3	0.751	0.763
Visual-CNN	84.5	0.812	0.825
Dual-BiLSTM	88.2	0.861	0.872
Fusion-AttendNet	94.7	0.932	0.944

Analysis indicated that the proposed model achieved an accuracy of 94.7%, improving the F1-score compared to single-input approaches and early fusion methods.

A cross-modal attention mechanism was integrated to learn adaptive weights between visual and environmental data to improve classifier stability, especially when the input was ambiguous, like at the transition state from attunement to drowsiness.

4.3. Environmental Comfort Prediction

Regression-based architecture enables the model to predict a single environmental comfort score through processing sensor data with occupancy information.

The evaluation of performance relied on Root Mean Square Error (RMSE) combined with Mean Absolute Error (MAE).

Table 4. Performance analysis of environmental comfort prediction

Model	RMSE	MAE
Env-LSTM	0.233	0.182
Dual-BiLSTM	0.197	0.151
Proposed Model	0.121	0.091

The proposed integration outperformed baseline predictive methods across all assessment metrics regarding environmental comfort determination. The model provides enhanced performance because it engages contextual occupancy information and engagement metrics to determine thermal and air-quality indices.

4.4. Actuator Control Prediction

Real-time conditions serve as inputs for predicting the activation state of four devices, which include fans and lights, relays, and purifiers. The evaluator measured performance by using micro-averaged precision together with latency metrics.

Table 5. Performance analysis of actuator control prediction

Model	Precision (%)	Latency (ms)
Rule-Based	83.1	5.3
Dual-BiLSTM	90.7	12.6
Proposed Model	96.2	8.4

Despite delivering short response times, the rule-based baseline operated without understanding context, and its sensor noise pollution triggered inaccurate results. Real-time inference operations on the proposed model took less than ten milliseconds per decision while achieving better control performance.

4.5. Visualization and Insights

The proposed model tracked the evolution of user behavior over time successfully by detecting long-term changes effectively. The system gained knowledge to detect elevated drowsiness probabilities during times after lunch, accompanied by high CO_2 readings through Figure 2, while activating fan and air purifier processes accordingly.

Figure 3 reveals the attention weight distribution, which demonstrates that the model used facial input better at morning hours then transferred focus to environmental data while lighting was dim or objects blocked visibility. The adaptive weighting mechanism proves to be the fundamental benefit of the attention-based fusion layer.

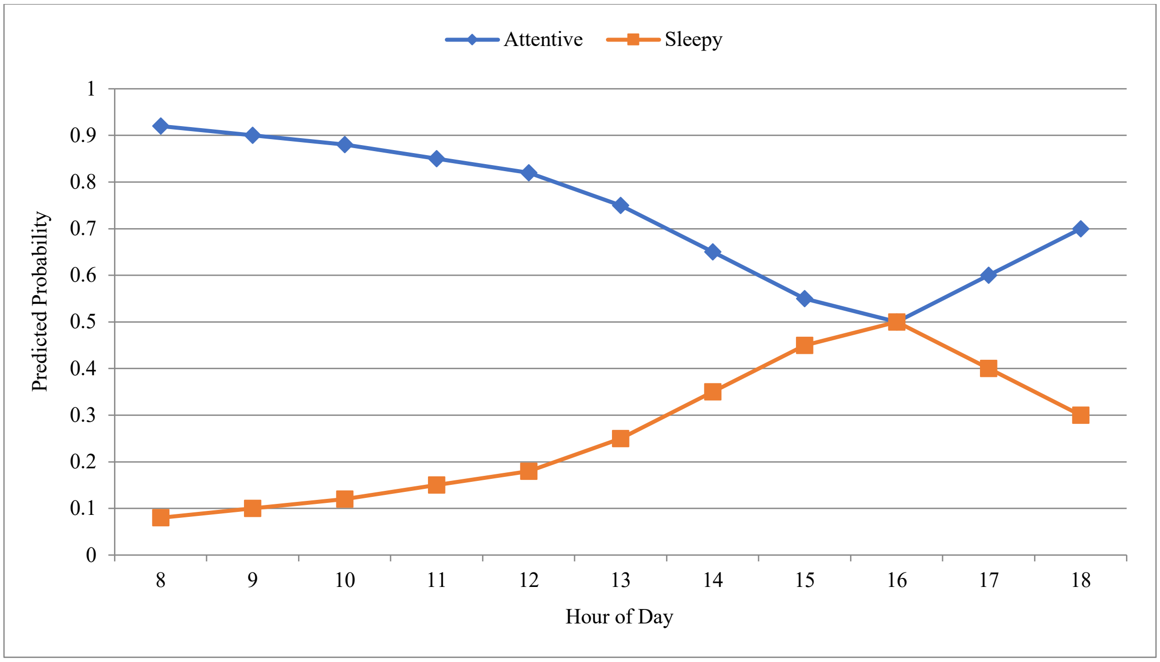

Fig. 2 Engagement state probability over a school day

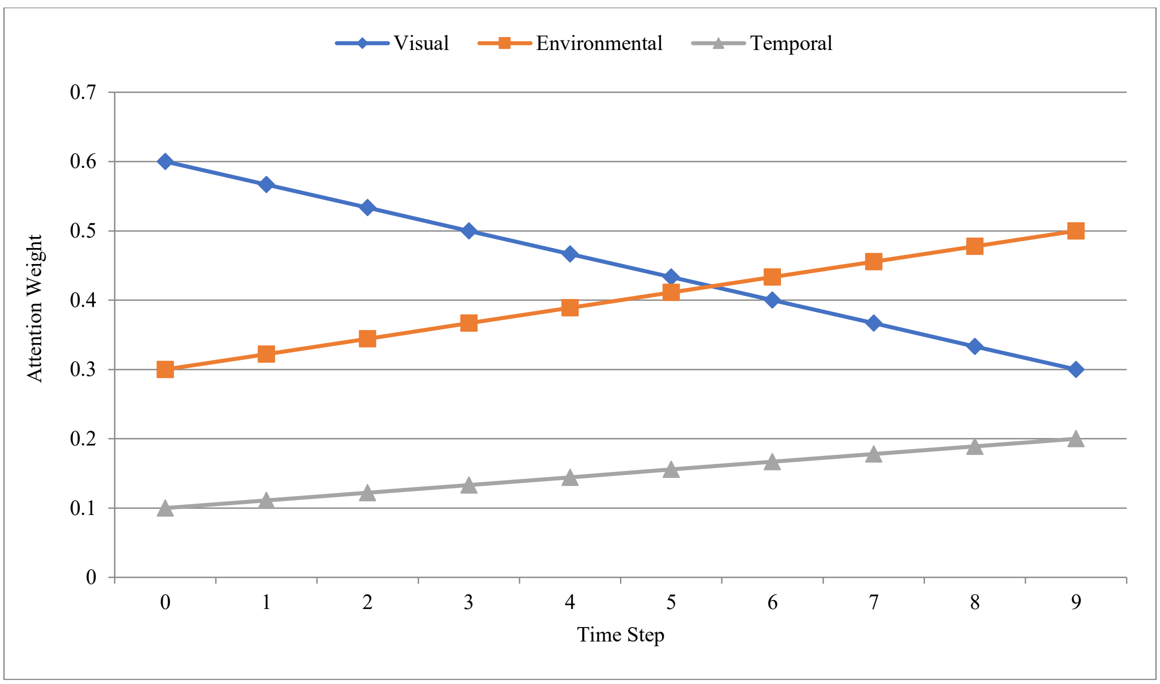

Fig. 3 Cross modal attention weight distribution over time

4.6. Generalization and Robustness

Although more statistical validation, like confidence intervals, per-class recall, and macro-averaged measures, show better performance, further statistical validation histories are to be conducted in long-term studies.

Moreover, the bias reduction was also performed by equalizing the class frequency in the design of the synthetic data to eliminate the skew towards the common behaviors. The training and validation loss kept on decreasing after 20 epochs, as demonstrated in Figure 4, meaning that there was a constant learning process with little overfitting. In the early stages, both validation and training losses were relatively high, with the validation loss being a little higher than the training loss, which is normally the norm in the initial stages of training because the weights and the model are not well optimized, and the model is also not certain. The training loss became nearly 0.36, and the validation loss became nearly

0.44 at the last epoch, showing good generalization with a consistent gap between the two curves. The presence of no drastic fluctuations proves the stability of the model and good regularization. The trends in accuracy, represented by Figure 5, also prove the success of the model in the classification of the engagement states. The accuracy of the training began at about 70 percent and gradually rose to approximately 96 percent at the 20th epoch. On the same note, the accuracy of the validation was also on an upward trend, with a close of approximately 94 percent at the close of training. The fact that the quality of training and validation were close to each other indicates that the model has successfully learnt the underlying patterns without overfitting. There are small variations in the validation curve (especially near epochs 5 and 11), which can be explained by batch-level variations, but do not point to any major instability.

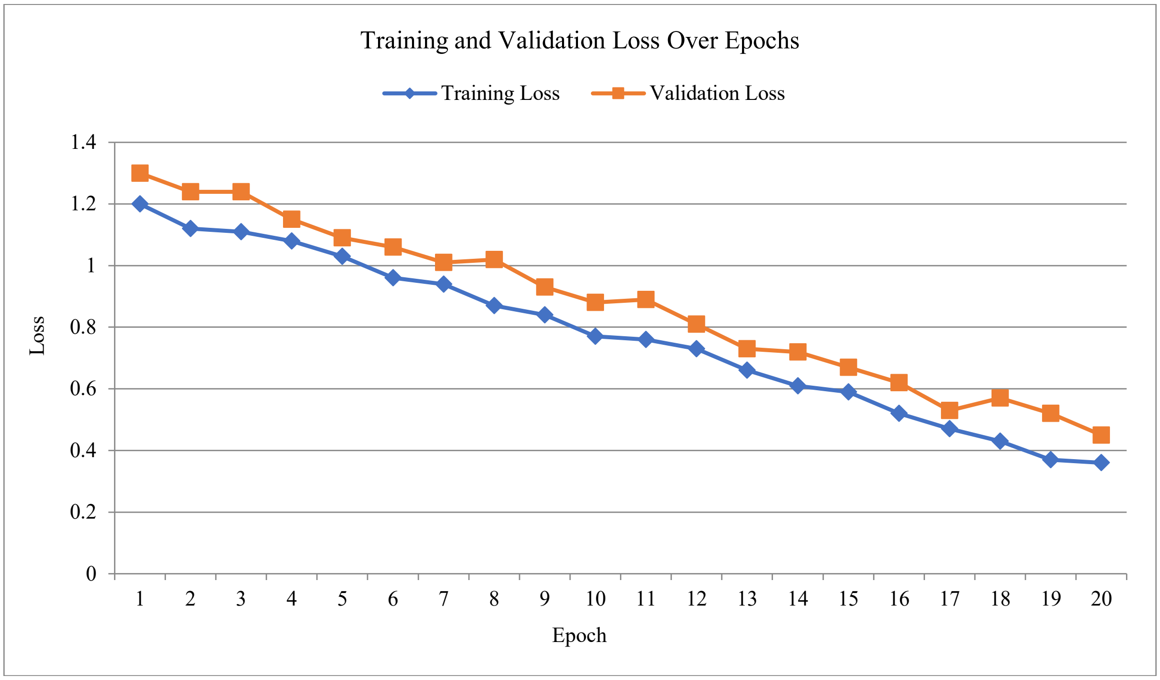

Fig. 4 Performance of loss of proposed model

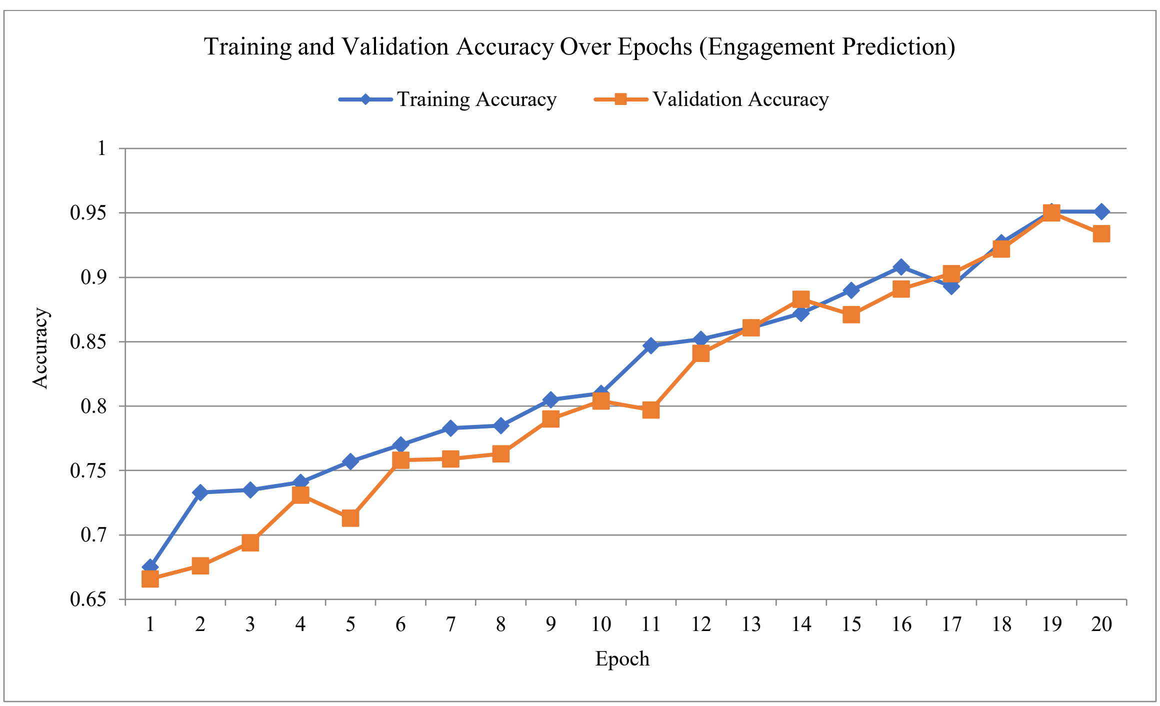

Fig. 5 Performance of accuracy of proposed model

4.7. Statistical Reliability and Confidence Interval Analysis

The statistical reliability analysis of the proposed Fusion-AttendNet model is shown in Table 6. The accuracy produced a standard deviation of 1.48 with a confidence error of ±0.84 and a p-value of 0.0031. It showed that the engagement classification performance was statistically stable. The RMSE had a low standard deviation of 0.016, a confidence error of ±0.008, and a p-value of 0.0064 for the prediction of environmental comfort. Likewise, the standard deviation of the MAE was 0.013, the confidence error was ±0.007, and the p-value was 0.0049, indicating a low prediction variation. Latency result: low SD (1.08), small confidence error (+/-0.66), small p-value (0.0038), which means that the latency real-time response behavior is consistent.

Table 6. Statistical reliability analysis of Fusion-AttendNet

Evaluation Component	Standard Deviation	Confidence Error	p-value
Accuracy	1.48	±0.84	0.0031
RMSE	0.016	±0.008	0.0064
MAE	0.013	±0.007	0.0049
Latency	1.08	±0.66	0.0038

5. Conclusion

The research developed Fusion-AttendNet, a new deep learning system that performs real-time engagement detection, environmental comfort evaluation, and control of smart classroom actuators. The proposed system merges environmental sensors with visual behavioral indicators and temporal data by using transformer technology together with cross-modal attention techniques, while diverging from existing isolated functional systems and unimodal information sources. The implemented system delivered 94.7% accuracy in engagement detection, alongside 0.121 RMSE for comfort assessment, along with 96.2% actuator control performance in less than 10 milliseconds. The experimental findings demonstrate that the combined attention mechanism and timing processing modules succeed at processing large-scale heterogeneous data fast enough for real-time systems.

Managed by the framework, the classroom devices, including fans and air monitors, operate autonomously to control physical learning conditions together with teaching advantages in educational spaces. The main innovative aspect of this study combines multiple data streams through a consolidated framework that performs simultaneous behavioral evaluation as well as environmental adjustments. Fusion-AttendNet advances classroom automation technology by linking the sensor to decision-making and to device control systems and offers scalable solutions to learning environments. The research analyzes future directions that involve adding audio signals and student feedback analytics to the system while deploying federated learning methods for decentralized and privacy-preserving model updates in various academic sites.

Acknowledgement

The current paper is a partial result of the project entitled "Developing and Implementing IoT Application in Higher Education Institutions: Exploring the Advantages and Challenges" funded by Arab Open University under the Internal Funding Program with ID: AOU_OM/2023/FBS2

References

[1] Divya Sharma et al., "Empowering Education: AI and IoT Integration for Smart Learning Environments," *Grenze International Journal of Engineering and Technology (GIJET)*, vol. 10, no. 2, 2024. [Google Scholar]

[2] Maria Efstratopoulou et al., "In their Own Voices: A Nationwide Study of Students Attitudes Towards the Implementation of Smart Learning Environments in UAE Schools," *Smart Learning Environments*, vol. 11, no. 1, pp. 1-22, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Kiran Budhrani, Yaeeun Ji, and Jae Hoon Lim, "Unpacking Conceptual Elements of Smart Learning in the Korean Scholarly Discourse," *Smart Learning Environments*, vol. 5, no. 1, pp. 1-26, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[4] Xieling Chen et al., "Past, Present, and Future of Smart Learning: A Topic-based Bibliometric Analysis," *International Journal of Educational Technology in High Education*, vol. 18, no. 1, pp. 1-29, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Adam Kenneth Dubé, and Run Wen, "Identification and Evaluation of Technology Trends in K-12 Education from 2011 to 2021," *Educational Information Technology*, vol. 27, no. 2, pp. 1929-1958, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Xiaomin Li, "Research on Intelligent Analysis Algorithm for Student Learning Behavior based on Multi-Source Sensor Data Fusion," *International Journal of High Speed Electronics and Systems*, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[7] Kashif Ishaq, and Samra Bibi, "IoT based Smart Attendance System using RFID: A Systematic Literature Review," *arXiv preprint*, pp. 1-26, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] M.T. Miranda et al., "Ventilation Conditions and their Influence on Thermal Comfort in Examination Classrooms in Times of COVID-19. A Case Study in A Spanish Area with Mediterranean Climate," *International Journal of Hygiene and Environmental Health*, vol. 240, pp. 1-15, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Sharmin Akter, and Asaduzzaman Asaduzzaman, "Internet of Things and Artificial Intelligence Enabled Smart Classroom," *2023 6th International Conference on Electrical Information and Communication Technology (EICT)*, Khulna, Bangladesh, pp. 1-6, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Abhinaya Saravanan et al., “EduVigil: Shaping the Future of Education with AI - An Intriguing Case Study,” *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India, pp.1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Tian-Tian Zhang, and Kai-Liang Zheng, “Learning Engagement through Multimodal Data Representation: A Theory-based Practical Framework,” *International Journal of Advances in Engineering and Management (IJAEM*), vol. 6, no. 11, pp. 232-241, 2024. [CrossRef] [Google Scholar]

[12] Aboul Hassane CISSE, “Real-time Adaptive Learning Environments using Gaze and Emotion Recognition Engagement and Learning Outcomes,” *International Conference on Computers in Education*, pp. 1-4, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Saad Razzaq et al., “DeepClassRooms: A Deep Learning based Digital Twin Framework for on-Campus Class Rooms,” *Neural Computing and Applications*, vol. 35, no. 11, pp. 8017-8026, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Pedro David Netto Silveira et al., “Smart Learning Environments in Support of Learning Ecosystems,” *Brazilian Symposium on Informatics in Education (SBIE). SBC*, pp. 175-185, 2022. [Google Scholar]

[15] Zhang Chao, Shi Qing, and Tong Mingwen, “A Study of Multimodal Intelligent Adaptive Learning System and its Pattern of Promoting Learners,” *Journal of Psychology Research*, vol. 13, no. 5, pp. 202-206, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Min Jae Lee, and Ruichuan Zhang, “Human-Centric Artificial Intelligence of Things–based Indoor Environment Quality Modeling Framework for Supporting Student Well-Being in Educational Facilities,” *Journal of Computing in Civil Engineering*, vol. 38, no. 2, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Kateryna Tanashchuk et al., “Adaptive Educational Ecosystems using Artificial Intelligence for Forming Students Educational Trajectories,” *Humanity Management*, vol. 25, no. 3, pp. 97-106, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Sucheta Yambal, and Yashwant Arjunrao Waykar, “AI-Driven Classroom Management: Balancing Efficiency, Engagement, and Ethics,” *Impacts of AI on Students and Teachers in Education 5.0. IGI Global Scientific Publishing*, pp. 93-124, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[19] Koutraki Maria, Efthymiou Vasilis, and Antoniou Grigoris, “S-CRETA: Smart Classroom Real-Time Assistance,” *Ambient Intelligence - Software and Applications: 3rd International Symposium on Ambient Intelligence (ISAmI 2012)*, Springer, Berlin, Heidelberg, vol. 153, pp. 67-74, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[20] Gyeonggeon Lee et al., “Multimodality of AI for Education: Towards Artificial General Intelligence,” *IEEE Transactions on Learning Technologies*, vol. 18, pp. 666-683, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[21] Stuart Hallifax et al., “Dynamic Gamification Adaptation Framework based on Engagement Detection through Learning Analytics,” *Companion Proceedings of the 11th International Conference on Learning Analytics and Knowledge LAK21*, pp. 1-14, 2021. [Google Scholar] [Publisher Link]

[22] Flavio Corradini et al., “REPTILE: A Proactive Real-Time Deep Reinforcement Learning Self-Adaptive Framework,” *arXiv preprint*, pp. 1-17, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[23] Deepti Chopra, and Praveen Arora, “Smart Classrooms, Smarter Teachers: A Deep Dive into AI-Driven Educational Enhancement,” *Journal of Informatics Education and Research*, vol. 3, no. 2, pp. 2813-2822, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[24] Hazel Farrell, *Re-Imagining Student Engagement in an AI-Enhanced Classroom*, 1st ed., Using Generative AI Effectively in Higher Education, Routledge, pp. 67-73, 2024. [Google Scholar] [Publisher Link]

[25] Deepti Chopra, and Praveen Arora, “Smart Classrooms, Smarter Teachers: A Deep Dive into AI-Driven Educational Enhancement,” *Journal of Informatics Education and Research*, vol. 3, no. 2, pp. 2813-2822, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[26] Aida Akavova, Zarema Temirkhanova, and Zarina Lorsanova, “Adaptive Learning and Artificial Intelligence in the Educational Space,” *E3S Web of Conferences*, *2nd International Conference on Environmental Sustainability Management and Green Technologies (ESMGT 2023)*, vol. 451, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[27] Lijuan Yan, Xiaotao Wu, and Yi Wang, “Student Engagement Assessment using Multimodal Deep Learning,” *Plos One*, vol. 20, no. 6, pp. 1-19, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[28] Jue Wang, Yuchen Sun, and Shasha Tian, “Deep Learning for Student Behavior Detection in Smart Classroom Environments,” *Information*, vol. 16, no. 11, pp. 1-23, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[29] Zhao Dan, Zhao Yali, and Zhao Tong, “A Multimodal AI Framework for Real-Time Student Engagement Detection and Adaptive Feedback in Higher Education,” *Journal of King Saud University Computer and Information Sciences*, vol. 38, no. 5, pp. 1-40, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[30] Kuan-Cheng Lin, Chiung-Chen Tseng, and Junyi Wu, “A Scalable Multimodal Framework for Learning Engagement Recognition using Three-Dimensional Convolutional Neural Networks and Semi-automatic Annotation,” *Frontiers in Artificial Intelligence*, vol. 9, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[31] Muddsair Sharif, and Dieter Uckelmann, “Multi-Modal LA in Personalized Education using Deep Reinforcement Learning based Approach,” *IEEE Access*, vol. 12, pp. 54049-54065, 2024. [CrossRef] [Google Scholar] [Publisher Link]